Extraction de motifs spatio-temporels : co-localisations, séquences et graphes dynamiques attribués

Frédéric Flouvat

Extraction de motifs spatio-temporels : co-localisations, séquences et graphes dynamiques attribués

Habilitation à Diriger des Recherches
Spécialité Informatique

Université de la Nouvelle-Calédonie

présentée et soutenue publiquement le 16 octobre 2019

par

Frédéric Flouvat

Rapporteurs :	Stéphane Bressan	Professeur Associé, National University of Singapore, Singapour
	Philippe Fournier-Viger	Professeur, Harbin Institue of Technology, Chine
	Alexandre Termier	Professeur, Université Rennes 1, France
Examinateurs :	Silvère Bonnabel	Professeur, Université de la Nouvelle-Calédonie, Nouvelle-Calédonie (président)
	Jean Diatta	Professeur, Université de La Réunion, France
	Nazha Selmaoui-Folcher	Maître de Conférences HDR, Université de la Nouvelle-Calédonie, Nouvelle-Calédonie (garant scientifique)

Institut de Sciences Exactes et Appliquées (ISEA) - EA 7484
Université de la Nouvelle-Calédonie
Campus de Nouville - BP R4 - 98851 Nouméa Cedex
https://isea.unc.nc

Mis en page avec la classe thesul.

Remerciements

Je souhaite remercier les membres du jury pour avoir bien voulu consacrer une partie de leur temps à mon HDR. Je remercie plus particulièrement Stéphane Bressan, Philippe Fournier-Viger et Alexandre Termier d'avoir accepté d'être mes rapporteurs. Leurs remarques et suggestions, tant sur le fond que sur la forme, ont été très pertinentes et appréciées.

Les travaux de recherche présentés dans ce mémoire sont le fruit de nombreuses collaborations. Je tiens donc à remercier l'ensemble des personnes qui ont favorisé grandement ces travaux. J'exprime ma gratitude à Nazha Selmaoui-Folcher pour m'avoir intégré dans ses projets dès mon arrivée et pour m'avoir permis de participer à des encadrements de thèses. Le travail présenté dans ce manuscrit est le résultat de cette collaboration. Je remercie les doctorants avec qui j'ai pu travailler et sans qui une grande partie des résultats n'auraient pas pu être obtenus : Hugo Alatrista Salas, Jérémy Sanhes, Zhi Cheng et Jannaï Tokotoko. Je souhaiterais aussi remercier les collègues avec qui j'ai pu collaborer à travers ces travaux (par ordre alphabétique) : Jean-François Boulicaut, Sandra Bringay, Antoine Collin, Dominique Gay, Claire Goiran, Laetitia Hédouin, Claude Pasquier et Maguelonne Teisseire. Je tiens à remercier tout particulièrement Jean-François Boulicaut pour ses conseils et sa disponibilité tout au long de ces années. Je remercie aussi Mohand-Saïd Hacid et Jean-Marc Petit pour leurs retours par rapport à cette HDR.

Plus généralement, je ne saurais oublier de remercier tous les collègues de l'ISEA pour leur soutien et pour les échanges enrichissant lors des pauses café et déjeuner. Le contexte fortement pluridisciplinaire dans lequel nous sommes n'est pas toujours simple, mais il est une vraie source d'enrichissement.

Je tiens à remercier chaleureusement mes amis pour leur présence pendant toutes ces années. Enfin et surtout, merci à Sylvie et Lily pour leurs attentions de chaque instant. Elles sont mes sources de joie, d'équilibre, et de motivation.

Table des matières

Partie I Introduction **1**

1 Contexte 3

2 Synthèse des travaux menés et organisation du manuscrit 5

Partie II Extraction de motifs spatio-temporels : état de l'art **9**

Chapitre 1 Données **11**

1.1 Spécificités des données spatio-temporelles 11

1.2 Principaux types de données 12

1.2.1 Les données liées à la mobilité 13

1.2.2 Les données d'évènements 14

1.2.3 Les données décrivant des informations continues sur des régions . . . 15

1.2.4 Les données de réseaux 17

Chapitre 2 Problèmes et principales approches **19**

2.1 L'extraction de motifs dans des séquences et des graphes 19

2.1.1 Les motifs séquentiels 19

2.1.2 Les graphes étiquetés, attribués et dynamiques 29

2.2 L'extraction de motifs dans des données spatio-temporelles 40

2.2.1 Suivi d'évènements ou d'objets spatiaux 40

2.2.2 Analyse d'une série temporelle de rasters 46

2.3 Positionnement des contributions 53

Partie III Contributions **57**

Chapitre 1 Extraction de co-localisations guidée par le domaine **59**

1.1 Cadre théorique 60

1.1.1 Domaine de motifs, contrainte et problématique 60

1.1.2 Parallèle avec la fouille d'*itemsets* 61
1.2 Intégration de contraintes spatiales et thématiques définies par les experts . . 62
1.2.1 Contraintes spatiales et thématiques 62
1.2.2 Intégration dans un algorithme d'extraction de motifs 63
1.3 Intégration de contraintes du domaine dérivées de modèles des experts 64
1.3.1 Les modèles des experts 65
1.3.2 Des motifs aux modèles 66
1.4 Visualisation cartographique des co-localisations 69
1.4.1 Comment représenter visuellement une co-localisation ? 69
1.4.2 Comment positionner une co-localisation sur une carte ? 70
1.5 Expérimentations et application à l'étude de l'érosion des sols 72
1.5.1 Prototype 72
1.5.2 Protocole expérimental 72
1.5.3 Analyse qualitative des motifs 73
1.5.4 Analyse quantitative 75

Chapitre 2 Extraction de motifs séquentiels intégrant le voisinage 77
2.1 Cadre théorique 78
2.1.1 Les données 78
2.1.2 Les motifs spatio-séquentiels 79
2.1.3 Les contraintes de fréquence et de participation minimales 80
2.1.4 Problématique 82
2.2 Stratégie d'extraction des motifs spatio-séquentiels 82
2.2.1 Algorithme par niveau dérivé d'*Apriori* 82
2.2.2 Algorithme en profondeur dérivé de *PrefixSpan* 83
2.3 Evaluation de la qualité des motifs extraits 85
2.4 Visualisation des motifs spatio-séquentiels 86
2.4.1 Représentation visuelle d'un motif spatio-séquentiel 87
2.4.2 Visualisation de l'ensemble des solutions 88
2.5 Expérimentations et application au suivi de la dengue 88
2.5.1 Prototype 88
2.5.2 Protocole expérimental 88
2.5.3 Analyse qualitative 89
2.5.4 Analyse quantitative 90

Chapitre 3 Recherche d'évolutions fréquentes en utilisant un graphe orienté acyclique attribué **93**

3.1 Cadre théorique . . . 95
3.1.1 Les données . . . 95
3.1.2 Les chemins pondérés . . . 96
3.1.3 Les contraintes de fréquence minimale et de non-redondance . . . 96
3.1.4 Problématique . . . 97
3.2 Stratégie d'extraction des chemins pondérés fréquents . . . 97
3.2.1 Extensions des motifs et projections du graphe . . . 97
3.2.2 Parcours en profondeur, structure de données et optimisations . . . 99
3.3 Intégration dans un processus d'analyse d'une série d'images satellitaires . . . 103
3.4 Expérimentations et application à l'étude de l'érosion des sols . . . 105
3.4.1 Prototype . . . 105
3.4.2 Protocole expérimental . . . 107
3.4.3 Analyse qualitative . . . 108
3.4.4 Analyse quantitative . . . 109

Chapitre 4 Extraction de motifs récurrents dans des graphes dynamiques attribués **113**

4.1 Cadre théorique . . . 114
4.1.1 Les données . . . 114
4.1.2 Les évolutions récurrentes . . . 114
4.1.3 Les contraintes : structure, temporalité, redondance et fréquence . . . 115
4.1.4 Problématique . . . 117
4.2 Stratégie d'extraction des motifs récurrents . . . 117
4.2.1 Intersections de graphes attribués et motifs de taille 1 . . . 117
4.2.2 Génération incrémentale des motifs . . . 119
4.3 Expérimentations et applications . . . 120
4.3.1 Prototype . . . 120
4.3.2 Protocole expérimental . . . 120
4.3.3 Analyse qualitative . . . 121
4.3.4 Analyse quantitative . . . 122

Partie IV Conclusion et perspectives **125**

Chapitre 1 Synthèse **127**

Chapitre 2 Perspectives **131**

Bibliographie **137**

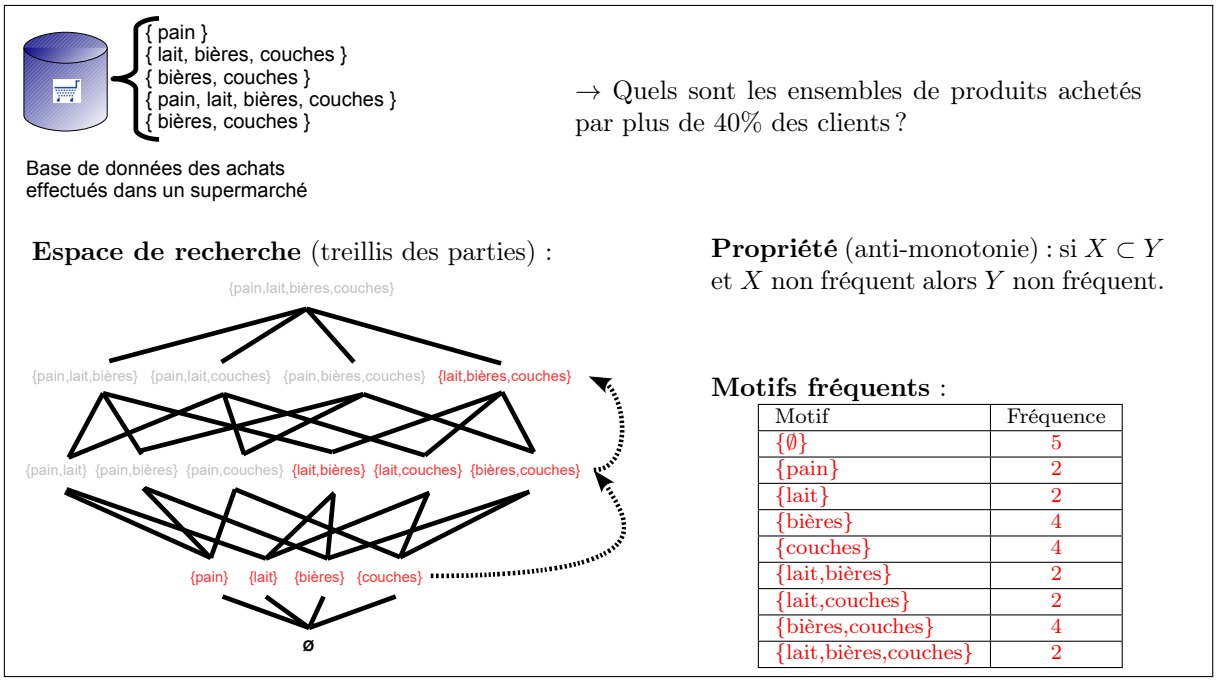
Première partie

Introduction

1 Contexte

Ces dix dernières années, la quantité de données collectées a explosé. Le commerce en ligne n'a jamais été aussi important, atteignant plus de 80 milliards d'euros en 2018 rien que pour la France. Tous les jours, les réseaux sociaux sont utilisés par des milliards de personnes. Les appareils connectés et les capteurs se sont démocratisés et ont investi les maisons et les entreprises. Au delà des volumes de données engendrés, leur complexité a aussi considérablement augmenté. Les données manipulées sont classiquement multi-dimensionnelles, multi-sources, hétérogènes, et bruitées. Elles sont de plus associées à des phénomènes caractérisés par des structures et des dynamiques complexes. Même si les avancées en matière d'analyse ont été importantes ces dernières années, les verrous scientifiques restent encore nombreux avant d'avoir une analyse intégrant toute cette complexité, tout en étant performante, robuste, pertinente et interprétable pour les experts du domaine. L'augmentation de la puissance de calcul des ordinateurs, bien que continue, n'est pas suffisante face à des tâches d'analyse de plus en plus complexes. Cette problématique est la raison principale de l'engouement récent autour du "*big data*" et du métier de "*data scientist*".

Dans ce contexte, une problématique a été plus particulièrement étudiée par la communauté : l'extraction de motifs intéressants (*pattern mining*). Ces motifs représentent des régularités suivies par une partie des données, i.e. des modèles locaux (par opposition aux modèles globaux construit par les approches de classification). Cette problématique a été introduite dans les années 90 avec pour application l'analyse des paniers d'achats (cf. exemple figure 1).



Motif	Fréquence
{∅}	5
{pain}	2
{lait}	2
{bières}	4
{couches}	4
{lait,bières}	2
{lait,couches}	2
{bières,couches}	4
{lait,bières,couches}	2

FIGURE 1 – Exemple d'extraction d'*itemsets* fréquents dans une base de données de transactions

La base de données en entrée était une collection de transactions effectuées par des clients. Chaque transaction était composée de plusieurs articles (ou *items*) achetés par un même client. L'objectif était alors d'extraire des ensembles d'articles (ou *itemsets*) fréquemment achetés ensemble. Un *itemset* est fréquent s'il apparaît dans plus d'un certain nombre de transactions (d'après un seuil *minsup* fixé par l'utilisateur). Un grand nombre de stratégies et d'algorithmes

ont été développés autour de cette problématique [AS+94, ZPO+97, Bay98, HPY00, BCG01, UAUA03] avec un objectif double : mettre en avant des propriétés permettant de limiter l'espace de recherche étudié et optimiser l'accès aux données. L'un des premiers algorithmes pour extraire ces motifs a été *Apriori* [AS+94]. Cet algorithme suit une stratégie "générer-tester". Des motifs candidats sont générés à partir des motifs solutions trouvés à l'itération précédente, puis leur fréquence est testée. L'espace de recherche (le treillis des parties) est parcouru "par niveau" (i.e. en largeur) et une propriété de la contrainte de fréquence ("l'anti-monotonie") est exploitée afin d'élaguer certains motifs candidats sans avoir à les tester. Une grande variété de types de motifs (domaines de motifs) a été étudiée depuis, tels que les séquences [AS95], les arbres [TRS02] ou les graphes [IWM00]. Toutefois, ce premier travail reste assez emblématique des différentes questions et des différents verrous à aborder lorsque l'on fait de l'extraction de motifs : la représentation des données, le domaine de motifs, les contraintes et l'algorithme d'extraction. Ils conditionnent l'information analysée, le type de motifs étudié, les motifs extraits et comment ils le sont.

Au delà de l'analyse de paniers d'achats, l'extraction des *itemsets* fréquents a aussi été utilisée dans des domaines variés tels que l'éducation, la santé, la cyber-sécurité, la justice ou le développement logiciel. Par exemple, [MBTP04] ont recherché des motifs et des règles mettant en avant les différences entre des groupes d'étudiants dans une université. Ils ont découvert différentes corrélations dont une montrant que les étudiants avec des notes moyennes au lycée, mais qui font leurs devoirs maison, ont 83% de chances de réussir. Dans [SK14], les auteurs font une étude comparant les résultats des garçons et des filles en mathématiques en Finlande, un pays qui affiche de bons résultats à ce niveau. Les motifs découverts montrent notamment que, malgré leur réussite, les étudiantes ont moins confiance en leurs compétences en mathématiques et ont moins l'intention de continuer dans ce domaine. La santé est aussi un domaine où les *itemsets* fréquents ont beaucoup été utilisés. Par exemple, [AZC01] utilisent ces motifs pour détecter des anomalies dans des mammographies. Les auteurs testent différentes techniques de fouille de données (réseaux de neurones et règles d'association) pour détecter et classer les anomalies dans les images. Ils montrent notamment que les *itemsets* utilisés pour construire les règles permettent d'obtenir de meilleurs résultats lorsque le jeu est relativement équilibré, tout en étant plus rapide à extraire. [BDF+03] utilisent quant à eux des travaux similaires (analyse formelle de concepts) pour prédire la toxicité de certaines molécules à partir d'ensembles de sous-structures chimiques. [SVN10] recherchent des motifs rares liés à des maladies cardio-vasculaires, et montrent notamment que les sujets présentant un certain allèle ont plus de risques d'avoir un syndrome métabolique (problèmes cardiovasculaires et diabètes). Dans un autre contexte, la découverte de ces motifs a permis de détecter des achats frauduleux par cartes bancaires [SVCS09]. Dans ce travail, les *itemsets* fréquents sont utilisés pour générer des règles d'association, et mettre par exemple en avant que les jeunes hommes sont plus affectés par ce type de fraudes. Dans [CVDVM16], les auteurs les utilisent quant à eux pour détecter des virus informatiques dans des téléphones mobiles. Grâce à cela, ils ont pu identifier une attaque ciblant les transactions bancaires de clients Coréens et Chinois en 2014. Des applications ont aussi été mises en avant dans les domaines de la justice et de la sécurité, par exemple pour mettre en avant des discriminations pour l'obtention de crédits [RPT10] ou pour analyser les crimes dans un quartier de Hong Kong [NCLY07]. Les *itemsets* fréquents ont également été appliqués pour analyser les erreurs et les pratiques lors du développement de logiciels [LZ05, SDC+13].

Ces dernières années, un grand nombre d'applications de cette famille de méthodes sont liées à des données spatiales et temporelles. En effet, on constate une explosion de la quantité d'informations spatiales générées, et de plus en plus un suivi dans le temps. Les domaines étudiés sont encore une fois très variés : étude des mouvements migratoires d'animaux [HPT12], analyse du déplacement des touristes [ZM11], suivi d'ouragans [LHW07], prévention de la criminalité dans

une ville [Cel15], analyse du trafic automobile [LZC+11], ou suivi de l'activité du soleil [AA16]. Même s'il s'agit d'extraction de motifs, le domaine de motifs utilisés dans ces travaux n'est généralement pas les *itemsets*, car leur structure ne suffit pas à capturer finement les interactions spatiales et temporelles, tout en intégrant les autres dimensions d'analyse. Pour cette raison, la communauté a défini et étudié des domaines de motifs de plus en plus riches, et donc de plus en plus complexes à extraire efficacement. La partie II du manuscrit introduira les principaux travaux étudiant cette problématique ainsi que leurs limites.

2 Synthèse des travaux menés et organisation du manuscrit

Dans le cadre de ma thèse, j'ai travaillé sur cette problématique de l'extraction d'*itemsets* fréquents avec pour objectif de développer des algorithmes (et des implémentations) génériques s'adaptant aux caractéristiques des données. Depuis mon arrivée en 2008 à l'Université de la Nouvelle-Calédonie, je m'intéresse plus particulièrement à l'extraction de motifs sous-contraintes dans des données spatio-temporelles complexes et à leur exploitation par les experts du domaine. Les domaines d'application étudiés ont été variés : érosion des sols, propagation d'une maladie vectorielle, déplacements de chauves-souris, activité aquacole, etc. Le suivi environnemental au sens large a été l'application moteur ayant guidé mes travaux. L'analyse et la gestion des risques pour l'environnement sont des problèmes majeurs en Nouvelle-Calédonie et dans le monde. La Nouvelle-Calédonie est un *"hotspot"* de la biodiversité mondiale où se côtoient un lagon classé au patrimoine mondial de l'UNESCO et une industrie minière d'envergure mondiale, le tout dans un climat tropical avec des événements extrêmes (p.ex. cyclones, tremblements de terre). Les enjeux en matière de suivi environnemental sont donc importants. Ces dernières années, une grande quantité de données a été collectée pour surveiller cet environnement. L'un des principaux défis à partir de ces données est de mieux comprendre et prévoir les différentes dynamiques. Cependant, les phénomènes sous-jacents et les données collectées sont complexes et variés (p.ex. images satellitaires, données terrain, données issues de modèles). Les objets à étudier sont nombreux, évoluent sur différentes échelles de temps et ne sont souvent pas clairement identifiés. De plus, leur évolution est liée à de nombreux paramètres en interaction. L'analyse de telles données est difficile et nécessite des méthodes d'analyse avancées.

Les contributions présentées dans ce manuscrit s'inscrivent dans ce contexte. Elles sont organisées en quatre chapitres :

— le premier chapitre présente une première contribution visant à extraire des motifs spatiaux (des co-localisations) plus pertinents et plus facilement interprétables par les experts. La contribution dans ce travail ne se situe ni au niveau du domaine de motifs, ni au niveau algorithmique. Elle est dans l'identification de contraintes (expertes ou dérivées de modèles experts) pouvant être intégrées efficacement dans les algorithmes d'extraction existants, puis dans la proposition d'une approche permettant de visualiser les solutions de manière plus intuitive pour les experts.

— Le deuxième chapitre présente le travail de thèse de Hugo Alatrista-Salas dans lequel nous avons intégré la dimension temporelle à l'analyse et proposé un nouveau domaine de motifs (les motifs spatio-séquentiels). Ces motifs permettent d'analyser l'évolution d'une zone (fixe) tout en considérant les événements ou objets à proximité. Un algorithme a été proposé pour extraire ces motifs et un prototype a été implémenté pour les visualiser. Une mesure d'intérêt a aussi été développée pour éliminer les motifs contradictoires.

— Le troisième chapitre introduit en grande partie le travail de thèse de Jérémy Sanhes. Il propose de modéliser des dynamiques spatio-temporelles plus complexes sous la forme d'un

unique graphe orienté acyclique attribué (appelé *a-DAG*). Un nouveau domaine de motifs et un algorithme sont ensuite proposés pour extraire efficacement des évolutions fréquentes dans ce type de graphes. Ces solutions ont été intégrées dans la plate-forme KNIME sous la forme de *plugins*, et elles ont été combinées en un processus permettant d'analyser une série d'images satellitaires.

— Le dernier chapitre décrit le travail de thèse de Zhi Cheng sur l'extraction de motifs récurrents dans un graphe dynamique attribué (un graphe évoluant dans le temps et dont les noeuds sont associés à plusieurs informations). Ces graphes permettent d'intégrer totalement les dimensions spatiales et temporelles, mais sont beaucoup plus complexes à analyser. Face à ce problème, ce travail introduit un domaine de motifs et un algorithme originaux basés sur des évolutions récurrentes de composantes connexes sous contraintes et des intersections de graphes.

Le tableau suivant résume tous ces travaux et les collaborations associées. Il sera repris dans chaque section et associé aux publications correspondantes.

Master/Thèse	**Co-encadrements**	**Projets et collaborations**	**Thématiques**
C. Grison (M2, 2009)	N. Selmaoui-Folcher		co-localisations, contraintes
H. Alatrista Salas (PhD, 2009-2012)	M. Teisseire, N. Selmaoui-Folcher et S. Bringay	Projet MOM dengue, Univ. Montpellier, Météo NC DASS, IPNC, IRD	motifs spatio-séquentiels, mesure d'intérêt, visualisation
E. Desmier (M1, 2009-2010)	N. Selmaoui-Folcher	CNRT petits bassins versants, Univ. Polynésie Française	co-localisation, visualisation, classification
L. Mabit (M2, 2010)	N. Selmaoui-Folcher	Projet MOM dengue, Univ. Montpellier, Météo NC DASS, IPNC, IRD	motifs séquentiels
C. Paul-Hus (M2, 2011)	N. Selmaoui-Folcher	CNRT petits bassins versants	modèles de risque d'érosion
J. Sanhes (PhD, 2011-2014)	N. Selmaoui-Folcher et J.-F. Boulicaut	ANR FOSTER INSA Lyon, CNRS Univ. Nice	a-DAG, chemins fréquents, images satellitaires
C. Mu (M2, 2014)	N. Selmaoui-Folcher	CNRT petits bassins versants	a-DAG, chemins fréquents, optimisation
M. Collin (M2, 2015)	N. Selmaoui-Folcher	ANR FOSTER	KNIME, a-DAG, images satellitaires
Z. Cheng (PhD, 2014-2018)	N. Selmaoui-Folcher	Ifremer	graphe dynamique attribué

TABLE 1 – Synthèse des encadrements, des projets et des collaborations en lien avec l'extraction de co-localisations guidée par le domaine

Comme le montrera la suite de ce manuscrit, bien que nos motivations premières étaient l'analyse de données spatio-temporelles, une attention particulière a été portée à la généricité des méthodes développées. Les deux dernières contributions liées à l'analyse de graphes ne sont pas limitées aux données spatio-temporelles, mais peuvent être utilisées pour analyser toutes données représentables sous la forme d'un graphe (p.ex. des réseaux sociaux, des molécules ou du code logiciel). Des expérimentations avec d'autres types d'applications (p.ex. réseaux de co-auteurs) ont donc aussi été faites afin de mettre en avant cette généricité.

Par ailleurs, certains travaux réalisés pendant cette période ne seront pas présentés car moins en rapport avec le thème de ce manuscrit. On notera notamment les travaux réalisés en collaboration avec Claude Pasquier (CNRS, Université de Nice) sur la fouille d'arbres et de graphes attribués.

Avant de présenter le détail de ces contributions, un état de l'art détaillé va être fait dans la partie qui suit. L'objectif de cet état de l'art est de positionner nos travaux mais aussi de donner une vision d'ensemble des points à traiter pour extraire des motifs dans des données complexes.

Deuxième partie

Extraction de motifs spatio-temporels : état de l'art

1

Données

Sommaire

1.1 Spécificités des données spatio-temporelles 11
1.2 Principaux types de données 12
1.2.1 Les données liées à la mobilité 13
1.2.2 Les données d'évènements 14
1.2.3 Les données décrivant des informations continues sur des régions . . 15
1.2.4 Les données de réseaux 17

Il existe une grande variété de données intégrant à la fois une dimension spatiale et temporelle. Ces données présentent toutefois des spécificités communes ayant un impact important sur l'analyse (p.ex. auto-corrélation, hétérogénéité, continuité, etc). Quatre types de données et d'applications ont été plus particulièrement étudiées : 1) les données liées à la mobilité (*mobility data*), 2) les données d'évènements (*event data*), 3) les données décrivant des informations continues sur des régions (*field data*) et 4) les données de réseau (*network data*).

Les sections suivantes présentent les spécificités des données spatio-temporelles, les principaux types de données étudiés, ainsi que les formalisations communément utilisées.

1.1 Spécificités des données spatio-temporelles

L'intégration des dimensions spatiales et temporelles a mis en avant de nouvelles possibilités en terme d'analyse, mais cela a aussi introduit de nouveaux défis propres à ce type de données.

D'après la première loi de la géographie de Tobler [Tob70], "tout interagit avec tout, mais deux objets proches ont plus de chances de le faire que deux objets éloignés". Autrement dit, des observations faites à deux localisations proches ont beaucoup plus de chances d'être corrélées. Des dépendances spatiales existent. Par exemple, la végétation observée dans une zone est liée à l'environnement directe de celle-ci. Cette auto-corrélation spatiale ne se limite pas aux localisations proches. Des dépendances à "longues distances" peuvent aussi exister. Par exemple, des phénomènes climatiques tels que El Niño ou La Niña peuvent avoir une influence sur la végétation de zones très éloignées. Ces dépendances doivent être prises en compte lors de l'extraction. Comme discuté dans [SJA+15, AKK18], les algorithmes prenant pour hypothèse une indépendance entre observations aboutissent souvent à des résultats non pertinents.

Les données spatio-temporelles sont aussi particulièrement hétérogènes. Les instances ne sont pas uniformément distribuées dans l'espace et le temps. Cette hypothèse communément prise

pour des données plus classiques (p.ex. transactionnelles) n'est donc pas valable pour des données spatio-temporelles. Toutes les instances n'appartiennent pas forcément à la même population. Elles ne sont pas réparties de la même façon et ne sont pas régies par les mêmes règles. Par exemple, un bus et un vélo ne vont pas évoluer de la même façon en fonction des zones géographiques et des périodes de la journée. De la même manière, l'évolution de la végétation va dépendre des régions mais aussi des saisons et des phénomènes inter-annuels (p.ex. El Niño). Différentes régions géographiques et périodes temporelles peuvent donc avoir des distributions différentes.

Les dimensions spatiales et temporelles ont d'autres spécificités. Premièrement, elles sont souvent prépondérantes dans l'analyse par rapport aux autres dimensions. En effet, l'objectif est généralement d'étudier l'évolution dans l'espace et dans le temps d'objets ou de phénomènes. Deuxièmement, elles sont par nature continues. Par exemple, les déplacements de véhicules sont généralement représentés sur une carte, i.e. un espace continu à deux dimensions. Cette particularité a un impact important sur les méthodes d'extraction. En effet, une grande partie d'entre elles considère des données discrètes (encore appelées nominales ou catégorielles). Un regroupement des valeurs (discrétisation ou *clustering*) est donc effectué (en pré-traitement ou pendant l'analyse). L'influence de cette étape sur les résultats est très importante. Par exemple, deux évènements pourront être perçus comme arrivant au même endroit et/ou au même moment (ou inversement) en fonction des regroupements effectués.

Les données spatio-temporelles sont aussi caractérisées par différents types d'attributs : des attributs non spatio-temporels, des attributs temporels et des attributs spatiaux. Les deux premiers types d'attributs sont associés à des valeurs numériques ou nominales (p.ex. intervalles ou catégories). Les relations entre les valeurs sont relativement simples et explicites (p.ex. relation d'ordre ou relation arithmétique). Les attributs spatiaux sont différents. Ils sont associés à des localisations, des zones, des périmètres ou des formes. Les relations sont variées et nécessitent souvent des calculs plus complexes [ES02]. Les relations peuvent être ensemblistes (p.ex. union, intersection, appartenance), topologiques (p.ex. contiguïté, couverture, croisement), métriques (p.ex. distance, surface, périmètre), directionnelles (p.ex. dessus, dessous), etc. Les méthodes d'extraction n'intègrent que très partiellement ces spécificités, et ceci malgré leur importance pour l'analyse. Par exemple, une zone urbaine est naturellement appréhendée à différentes échelles tels que le bloc, le quartier, ou la ville. Cette relation hiérarchique peut exister pour d'autres types de données (p.ex. données d'une entreprise). Toutefois, elle est plus importante dans le cadre des données spatio-temporelles car les dépendances spatiales et temporelles entre les observations peuvent beaucoup dépendre de l'échelle considérée.

1.2 Principaux types de données

Une grande variété de données spatio-temporelles existe en fonction des objets étudiés, des contraintes d'acquisition et des problématiques. En géographie, les données spatiales sont classiquement associées à trois modèles : les objets, les champs continus et les réseaux spatiaux. Ces modèles correspondent aux différentes catégories d'objets utilisés pour représenter numériquement les informations géographiques du monde réel. Ils constituent les principaux types manipulés dans les systèmes d'information géographique. L'intégration de la dimension temporelle enrichit encore ces modèles spatiaux et ouvre de nouvelles perspectives. Ces modèles spatio-temporels représentent les catégories de problèmes communément étudiés dans la littérature. Le tableau 1.1 présente les principaux types de données, les modèles spatiaux considérés ainsi que les modèles spatio-temporels étudiés.

Données	Modèle spatial		Modèle spatio-temporel	Description
Mobilité (*mobility data*)	Objets (points, lignes, polygones)		ensemble de trajectoires	trajectoires d'objets mobiles suivis dans l'espace et le temps
Évènements (*event data*)			ensemble de séries temporelles spatiales	objets/évènements suivi continuellement dans le temps (uniquement)
			ensemble de types d'objets spatio-temporels	objets/évènements discrets arrivant ponctuellement dans l'espace et le temps
Région (*field data*)	Champ continu	espace régulier (raster ou grille de points)	série temporelle de champs continus	observations continues distribuées dans un espace (2D ou 3D) et changeant au cours du temps
		espace irrégulier (points, lignes, polygones)		
Réseau (*network data*)	Réseau spatial (graphe)		graphe dynamique, graphe temporel, réseau de flot	réseau d'objets localisés, associés à des métriques, évoluant au cours du temps (structure et métriques)

TABLE 1.1 – Les principaux types de données et leurs caractéristiques

Avant de renter dans le détail de chacun de ces types de données, il est intéressant de noter que la majeure partie d'entre eux représente l'espace de manière discrète. Les données "mobilité", "évènements" et "réseau" correspondent à des observations sur des objets et des évènements discrets dans l'espace. Seuls certains modèles de champs continus peuvent représenter l'espace de manière continue. Par exemple, une région peut être représentée par un espace en deux dimensions (p.ex. une carte) composé d'une juxtaposition de polygones irréguliers (p.ex. des parcelles agricoles, des forêts, ou des habitations). Même si certaines données sont issues d'un suivi "en continu" d'objets, le temps est considéré de manière discrète. En effet, les observations et mesures sont généralement enregistrées ponctuellement dans le temps.

1.2.1 Les données liées à la mobilité

L'émergence des nouvelles technologies mobiles (démocratisation des GPS et des téléphones mobiles) a entrainé la collecte de grandes quantités de données liées à la mobilité (*mobility data*). Ces données ont permis d'entrevoir de nouvelles applications dans des domaines très variés : de l'écologie (p.ex. migration d'oiseaux) au sport (p.ex. football), en passant par la météorologie (p.ex. suivi des ouragans) ou le tourisme [MCK+04, CMC05, GNPP07, LHJ+11, LZC+11, Pha13, HPL15, WDL+17]. A titre d'exemple, le projet européen GeoPKDD [GP08] a utilisé les données de déplacements des véhicules pour revoir l'aménagement du plan de circulation de grandes agglomérations.

Ces données relatives aux déplacements, et à la mobilité en général, se présentent généralement sous la forme de bases de données de trajectoires. Les trajectoires sont définies comme des objets en mouvement représentés par des séquences de tuples (id, l, t), où l est la localisation de l'objet id au temps t. La figure 1.1 représente schématiquement un ensemble de trajectoires d'oiseaux.

Dans certains cas, d'autres informations sont collectées sur les objets ou sur leur environnement. Par exemple, les déplacements des touristes dans une ville peuvent être rapprochés des monuments historiques ou des musées à proximité. De même, il est possible de suivre les ca-

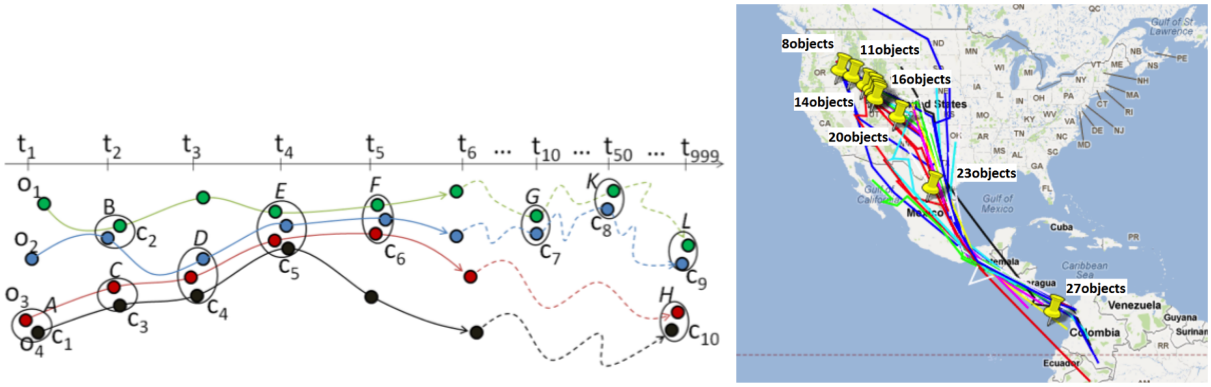


FIGURE 1.1 – Exemple de trajectoires d'oiseaux [Pha13]

ractéristiques d'un ouragan en même temps que l'on suit son déplacement. Ce croisement entre trajectoires, données géographiques et caractéristiques des objets est à l'origine du concept de "trajectoires sémantiques" [ABK+07, YCP+13, BRdA+14]. Dans ce modèle plus général, les trajectoires peuvent être représentées par une séquence de tuples (id, sp, t, ma, tag) où id est l'identifiant de l'objet suivi, sp est sa position sémantique (un objet spatial représentant la localisation et une information sémantique la décrivant), t est le temps, ma est une annotation sur le mouvement (p.ex. à l'arrêt ou en mouvement) et tag est un ensemble d'informations associé à l'objet suivi au temps t.

1.2.2 Les données d'évènements

Des données décrivant des événements (*event data*) sont également acquises dans une grande variété d'applications tels que l'épidémiologie (p.ex. propagation d'une épidémie), l'écologie (p.ex. feux de forêts), les transports (p.ex. accidents de la route), la criminologie (p.ex. crimes dans une ville), et les réseaux sociaux (p.ex. tweets) [HZZ08, CSRS08, DLSL09, ASG13, FK14, Yu16]. Classiquement, ces données correspondent à l'historique d'évènements arrivés à un endroit et à un instant donné. Par exemple, un accident de voiture peut être caractérisé par sa nature, sa localisation ainsi que la date à laquelle il est arrivé. Un évènement peut ainsi être décrit par un tuple (l, t, c) où l est la localisation de l'évènement de type c au temps t. La figure 1.2 présente cinq types d'évènements (A, B, C, D et E) arrivant à trois temps successifs (t_1, t_2 et t_3). Cette figure représente les évènements par des points dans un espace euclidien à deux dimensions.
Contrairement aux trajectoires, ces données contiennent plusieurs types d'objets et d'évènements à étudier en simultané. De plus, les objets/évènements ne sont pas identifiés de manière unique et les pas de temps ne sont généralement pas réguliers.

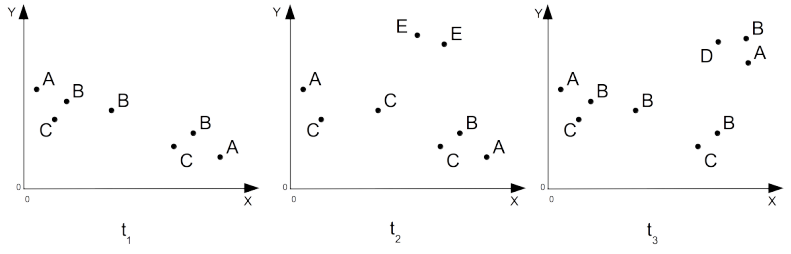


FIGURE 1.2 – Cinq types d'évènements arrivant à trois temps successifs

Toutes les données liées à des évènements ne se limitent pas nécessairement à cette définition. Bien que les évènements soient souvent associés à des points dans l'espace, ils peuvent aussi être décrits par des polygones. Par exemple, un feu peut être représenté par un polygone délimitant la zone impactée. Dans certains cas, d'autres informations sur l'évènement peuvent aussi être connues. Au delà du type d'accident, cet évènement peut par exemple être caractérisé par le nombre de véhicules impliqués, l'âge des conducteurs ou le nombre de blessés. Ces données peuvent aussi contenir des informations sur les autres objets/évènements présents au même moment que l'objet d'étude. Dans le cadre d'une ville, on peut par exemple avoir les bâtiments, les évènements sportifs au même moment, etc. Les évènements peuvent aussi avoir une durée ou une date de début et de fin. Par ailleurs, même si les évènements sont souvent représentés dans un espace euclidien, il peut être intéressant pour certaines applications de les représenter autrement. Par exemple, la distance euclidienne n'est pas nécessairement la meilleure distance pour mesurer l'éloignement entre deux accidents. Une distance prenant en compte le réseau routier peut être plus pertinente.

Dans le cas le plus général, une base de données d'évènements est donc un ensemble de types d'objets (*object-type*) spatio-temporels. Chaque occurence d'un type d'objets (ou évènements) est caractérisé par un tuple (l, t, c, o, p) où l est la localisation de l'objet spatial o (point, ligne ou polygone) de type c, et associé à l'ensemble de propriétés p, au temps t. Lorsque les objets/évènements ont une localisation fixe, ces données constituent des séries temporelles spatiales et peuvent être représentées par des séquences de valeurs associées à une localisation.

1.2.3 Les données décrivant des informations continues sur des régions

Les entités géographiques sont classiquement regroupées selon deux principaux modèles : les objets discrets et les champs continus (*continuous field*) [GG89, WD04, GYC07, BMML15]. Le premier type voit l'espace comme un ensemble d'objets décrits par des attributs et localisés en fonction d'un sytème de coordonnées géométriques. Le deuxième type décrit la distribution d'attributs variant de manière continue dans une région de l'espace. Comme indiqué dans [Par94], "*un champ continu est une portion de l'espace où la force appliquée à un point dépend de sa position seule. Ces champs sont appelés 'continuous fields', appellation qui renforce la notion de continuité des valeurs des champs : le nombre de points contenus dans un champ est infini, la représentation des valeurs de ce champ est donc continue.*" Un champ continu peut correspondre à une valeur (numérique ou nominale) ou à un vecteur de valeurs en fonction du nombre de mesures associé à chaque point. Les propriétés d'un environnement (p.ex. la pollution, la végétation ou la température) sont naturellement perçues comme des champs continus variant en fonction de la localisation. L'activité dans un cerveau peut aussi être vue comme un champ continu.

Les ordinateurs stockent les informations de manière discrète sur des supports avec des capacités limitées. De plus, l'acquisition de valeurs en continu dans l'espace peut être difficile, voire impossible, dans certains cas. Ces champs continus sont donc nécessairement échantillonnés et discrétisés. Dans ce contexte, différents modèles ont été définis pour représenter ces champs continus (cf.figure 1.3). Ils diffèrent dans leur façon de représenter l'espace et dans la complétude des informations stockées. Tout d'abord, l'espace peut être représenté de manière régulière ou irrégulière. Comme le montre la figure 1.3, les rasters et les grilles de points découpent l'espace de manière régulière. Les pavages de polygones, les courbes de niveau, les réseaux de triangles irréguliers (TIN en anglais) et les points irrégulièrement espacés découpent quant à eux l'espace de manière irrégulière. Ensuite, le champ peut être représenté de manière complète ou incomplète. Les rasters, les pavages de polygones, les courbes de niveau, et les réseaux de triangles irréguliers sont des représentations complètes car le champ continu est estimé en chaque point

d'une région. De plus, les régions recouvrent complètement l'espace étudié. Les grilles de points et les points irrégulièrement espacés sont des représentations incomplètes car le champ continu est uniquement défini en certains points. Dans ce cas, une fonction mathématique doit être utilisée pour estimer le champ à des localisations non échantillonnées.

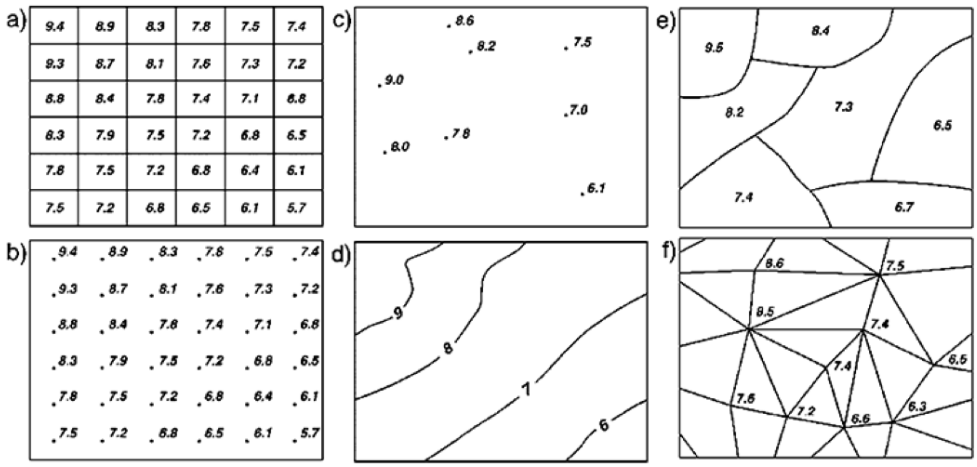


FIGURE 1.3 – Différentes représentations d'un champ continu [PSTV10] : a) raster b) grille de points c) points irrégulièrement espacés d) courbe de niveau e) pavage de polygones f) réseau triangulé irrégulier (TIN)

Le raster est probablement le modèle le plus utilisé et étudié, car il est généralement associé aux images numériques. Ces dernières années, l'utilisation de données rasters s'est multipliée dans des domaines telles que la médecine, l'astronomie, la météorologie ou l'écologie. Les avancées technologiques en imagerie satellitaire ont par exemple permis d'avoir des données plus précises, plus riches en informations et couvrant des zones et des temps plus importants, pour un coût moins élevé. Elles ont ainsi ouvert de nouvelles perspectives en terme de suivi et mis en avant de nouveaux défis [MWW$^+$15].

Dans ce modèle, l'espace est découpé selon une grille composée de cellules rectangulaires (les pixels). Chaque pixel correspond à un vecteur de valeurs $(r_1, r_2, ...r_K)$ où K est le nombre d'attributs du champ continu (appelé aussi nombre de bandes de l'image). Par exemple, un pixel rouge d'une image RGB aura pour valeur $(255, 0, 0)$. Si le raster représente la température et le taux d'humidité d'une région, un pixel associé à 27° et 80% d'humidité aura pour valeur $(27, 0.8)$. La taille des cellules de la grille par rapport à la région représentée en réalité constitue la résolution spatiale du raster. Par exemple, la résolution du raster sera de 50 cm, si chaque pixel représente en réalité une région rectangulaire de 50×50 cm. Cette résolution dépend principalement du matériel d'acquisition ou de la compression souhaitée. Plus formellement, un raster peut être défini par un tenseur, i.e. un tableau multidimensionnel de valeurs. Une image satellitaire sera par exemple associée à un tenseur $R \in \mathbb{R}^{I,J,K}$ où I et J sont les dimensions de l'image en pixels, et K est son nombre de bandes. Un élément du tenseur $r_{i,j,k}$ représente la valeur du k-ième attribut du champ continu pour le pixel (i, j).

L'évolution d'une région par rapport à un champ continu peut être étudiée en enregistrant dans différents rasters les informations du champ continu au cours du temps. La série temporelle de rasters ainsi constituée peut être définie par une séquence de tuples (R, t) où R est le tenseur associé au raster au temps t. La figure 1.4 illustre cela pour une série de rasters sur trois temps consécutifs.

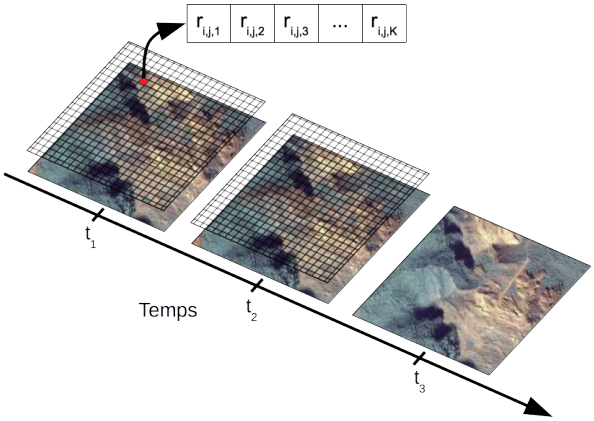


FIGURE 1.4 – Série temporelle de trois rasters

1.2.4 Les données de réseaux

Le concept de réseau apparaît dans de nombreux domaines d'applications. On le retrouve par exemple dans les transports (p.ex. réseau routier), dans les communications (p.ex. internet), en environnement (p.ex. réseau hydrologique) et en biologie (p.ex. réseau neuronal). De manière très générale, on pourrait définir un réseau comme un ensemble d'entités interconnectées par des liens de nature diverse. Par exemple, la figure 1.5 représente le réseau ferroviaire d'Auckland, i.e. différentes localisations connectées entre elles par des lignes de trains. De par leur importance, ces réseaux sont au centre de beaucoup d'attentions depuis de nombreuses années. Les entités les composant, tout comme les liens entre celles-ci, font souvent l'objet d'un suivi et donc d'une collecte de données. Tout comme pour les autres types de données spatio-temporelles, cette collecte est de plus en plus importante en raison de la multiplication des capteurs et de l'augmentation de leur précision.

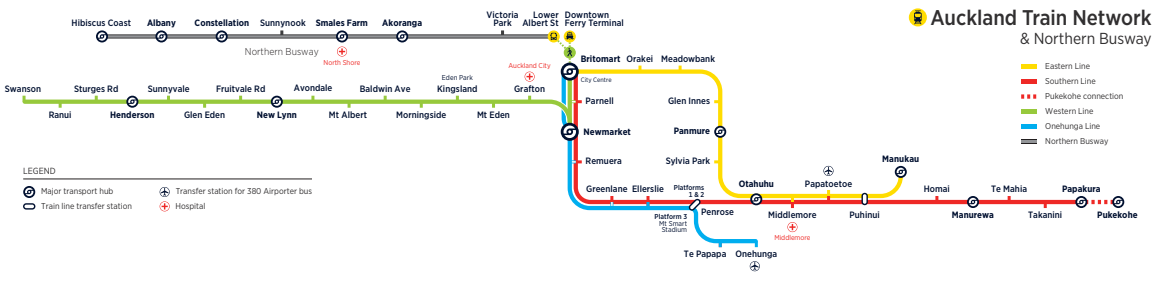


FIGURE 1.5 – Carte du réseau ferroviaire d'Auckland (*https ://at.govt.nz*)

Dans beaucoup de ces infrastructures en réseau, la dimension spatiale est particulièrement importante. Par exemple, les réseaux de transports sont de manière évidente représentés dans l'espace et liés à une notion de distance, tout comme certains réseaux de communications (p.ex. wifi). Comme indiqué dans [Bar11], les réseaux spatiaux (*spatial networks*) sont des réseaux dans lesquels les entités sont localisées dans l'espace et associées à des métriques. Ces réseaux ont des contraintes géométriques et ont donc des propriétés topologiques particulières. Souvent, les réseaux sont représentés dans un espace à deux dimensions et étudiés en fonction de la distance euclidienne. Dans ces cas, la probabilité d'avoir une connection entre deux entités décroit quand la distance augmente. Ces réseaux peuvent être planaires (p.ex. réseau ferroviaire) ou non (p.ex.

réseau aérien). De plus, les liens entre les entités peuvent ne pas être spatiaux. Par exemple, les liens entre des personnes d'un réseau social ne sont pas spatiaux, alors que les personnes sont associées à une localisation (voire plusieurs). Toutefois, l'espace intervient de manière indirecte, car la probabilité que deux personnes soient "amies" est plus forte lorsqu'il y a proximité spatiale.

Classiquement, un réseau spatial est représenté sous la forme d'un graphe $G = (V, E, \lambda_V, \lambda_E)$, où V est l'ensemble des noeuds (localisations), E est l'ensemble des arêtes (liens), et λ_V (resp. λ_E) est une fonction d'étiquetage qui associe un noeud (resp. une arête) à un ensemble d'informations (numériques ou nominales). Par exemple, le réseau ferroviaire illustré en figure 1.5 peut être représenté sous la forme d'un graphe où les noeuds correspondent aux stations et les arêtes aux lignes de trains. Les noeuds sont associés à des informations liées à l'environnement spatial de la station (p.ex. aéroport ou hôpital) ou à son type (p.ex. *hub*). Les arêtes sont associées à des durées de transport, des nombres de passagers, ou des types de lignes ferroviaires (p.ex. train rapide). Lorsque le réseau évolue dans le temps, on obtient un graphe dynamique, i.e. un graphe dont les noeuds, les arêtes et/ou les informations associées peuvent changer au cours du temps. Plus formellement, un graphe dynamique G peut être représenté par une séquence $\langle G_{t_1}, G_{t_2}, ..., G_{t_{max}} \rangle$, où chaque graphe G_t représente le graphe G au temps $t \in \{t_1, \ldots, t_{max}\}$.

2

Problèmes et principales approches

Sommaire

2.1 L'extraction de motifs dans des séquences et des graphes 19
2.1.1 Les motifs séquentiels 19
2.1.2 Les graphes étiquetés, attribués et dynamiques 29
2.2 L'extraction de motifs dans des données spatio-temporelles . . 40
2.2.1 Suivi d'évènements ou d'objets spatiaux 40
2.2.2 Analyse d'une série temporelle de rasters 46
2.3 Positionnement des contributions 53

Le chapitre précédent a mis en avant la diversité et la complexité des données spatio-temporelles collectées, ainsi que leur intérêt dans un grand nombre de domaines. Face à ces données, les besoins en matière d'analyse et d'aide à la décision sont importants. Dans la suite, ce manuscrit va se focaliser plus particulièrement sur une de ces tâches : l'extraction de motifs, de régularités, cachés dans de telles données.

Beaucoup de travaux ont étudié l'extraction de motifs intéressants dans des données spatio-temporelles. En général, ils se concentrent sur l'étude de domaines de motifs, de contraintes, ou de pré/post traitements propres à ce type de données (p.ex. détection d'objets spatiaux, contraintes de voisinage, etc.). Toutefois, les motifs solutions sont généralement générés par des stratégies ou des algorithmes développés dans des contextes plus généraux. Les travaux effectués dans le cadre de la fouille des séquences et de graphes ont notamment été beaucoup utilisés en raison de l'intérêt de ces domaines de motifs pour représenter les dimensions temporelles et spatiales. Ce chapitre commencera donc par présenter ces travaux, puis il montrera comment ceux-ci ont été adaptés dans le cadre de données spatio-temporelles. La suite de ce manuscrit va plus particulièrement se focaliser sur l'extraction de motifs spatio-temporels dans des données géographiques de types événements et rasters.

2.1 L'extraction de motifs dans des séquences et des graphes

2.1.1 Les motifs séquentiels

L'extraction de motifs séquentiels est une des problématiques les plus étudiées en fouille de données. Elle a des applications multiples telles que l'analyse d'achats en ligne [SA96b], l'analyse d'exécutions de logiciels [LKL08], l'analyse des usages du Web [ME10], ou l'étude des communications mobiles [YZC12].

La donnée en entrée est une collection de séquences. Chaque séquence correspond à une "transaction" et elle est composée d'une succession d'ensembles de valeurs catégorielles (des *itemsets*). Le tableau 2.1 présente un exemple de base de données séquentielles. L'objectif est d'extraire des sous-séquences intéressantes dans ces données. Une des contraintes régulièrement utilisée pour déterminer l'intérêt d'un motif est la fréquence minimale. La mesure de fréquence est généralement définie comme le nombre de transactions contenant la sous-séquence (le motif). Cette fréquence est donc différente du nombre d'occurrences du motif car celui-ci peut apparaître plusieurs fois dans une même transaction. Toutefois, calculer le nombre total d'occurrences est très coûteux en raison de la multitude d'entrelacements possibles. De plus, la contrainte qui en découle peut ne pas être (anti-)monotone, ce qui limite les possibilités d'élagage de l'espace de recherche lors de l'extraction (et donc le passage à l'échelle des algorithmes).

SID	**Séquence**
1	$< \{a, c\}, \{a, b\}, \{c, f\}, \{e, f\} >$
2	$< \{a, c\}, \{a, c\}, \{d, f\} >$
3	$< \{a, c\}, \{a, c, e\}, \{c, e, f\}, \{f\} >$
4	$< \{c\}, \{a, d\}, \{b, e\}, \{e\} >$
5	$< \{b, d\}, \{c, f\} >$

TABLE 2.1 – Base de données de séquences

Extraire des sous-séquences fréquentes : principales stratégies Un grand nombre d'algorithmes et de contraintes ont été proposés pour extraire des motifs séquentiels. Ils utilisent tous des stratégies différentes mais sont généralement basés sur les deux mêmes opérations pour construire de nouveaux motifs : l'extension du dernier *itemset* de la séquence, ou l'extension de la séquence, par un *item*. Par exemple, la séquence $< \{a, c\} >$ peut être étendue par l'*item* e de deux façons : $< \{a, c, e\} >$ et $< \{a, c\}, \{e\} >$. La suite de cette section présentera les spécificités des principales approches de la littérature. Des états de l'art plus détaillés peuvent être trouvés dans [ME10, MR13, FVLK$^+$17].

AprioriAll est l'un des premiers algorithmes développé pour extraire des sous-séquences fréquentes [AS95]. Il s'appuie sur la même stratégie que l'algorithme *Apriori* [AS$^+$94] des mêmes auteurs. Il effectue un parcours en largeur de l'espace de recherche (approche par niveau) et s'appuie sur une stratégie de type "générer-tester". L'algorithme utilise les motifs fréquents de taille k (i.e. contenant k *items*) pour générer les motifs candidats de taille $k+1$. Puis, il élague ceux ayant un sous-motif de taille $k-1$ non fréquent, et vérifie la fréquence des motifs restants. La base de données est stockée en mémoire sous un format relativement similaire à celui du tableau 2.1 (base de données horizontale). Les auteurs étendent leurs travaux dans [SA96b] et proposent un algorithme plus générique (appelé GSP) intégrant des contraintes temporelles (fenêtre glissante et intervalle entre deux évènements). Leur approche prend aussi en compte des taxonomies définies par les utilisateurs, ce qui permet d'extraire des motifs à différentes échelles. Une structure de données de type arbre de hachage (*hash-tree*) est aussi utilisée pour réduire le nombre de motifs candidats à tester et améliorer l'efficacité du comptage de la fréquence. Dans [MCP98], les auteurs adoptent la même stratégie mais utilisent un arbre préfixe (*prefix-tree*) pour stocker les motifs candidats et ainsi améliorer l'efficacité de l'approche (mémoire consommée et nombre de traitements effectués). L'algorithme $SPIRIT$ développé dans [GRS99] permet d'intégrer, en plus de la fréquence minimale, une contrainte $\mathcal{C}$ basée sur une expression régulière définie par les utilisateurs. Leur algorithme est similaire à l'algorithme GSP [SA96b], i.e. parcours par niveau

de l'espace de recherche basé sur la même stratégie "générer-tester". La principale difficulté réside dans l'intégration de la contrainte $\mathcal{C}$ et son impact sur l'efficacité de ce type de stratégie. L'ajout de la contrainte $\mathcal{C}$ augmente l'efficacité de la génération des candidats et de l'élagage lorsque celle-ci est anti-monotone. Toutefois, elle peut en diminuer l'efficacité dans le cas contraire. En effet, dans ce cas, une séquence de taille k ne vérifiant pas $\mathcal{C}$ ne peut plus être utilisée pour élaguer ses sur-séquences de taille $k+1$. Pour autant, la contrainte $\mathcal{C}$ va diminuer le nombre de motifs candidats générés et donc le nombre potentiel de motifs non fréquents pouvant être utilisés pour élaguer l'espace de recherche. De plus, des séquences de taille $k+1$ peuvent être solutions sans qu'aucune des sous-séquences de taille k ne vérifie $\mathcal{C}$, ce qui pose des problèmes pour la génération des candidats. Face à ces problèmes, l'une des solutions est de trouver une relaxation anti-monotone de la contrainte $\mathcal{C}$. Toutefois, il n'en existe pas toujours une. Les auteurs proposent donc quatre algorithmes intégrant différents degrés de relaxation de $\mathcal{C}$. Ils utilisent l'automate fini déterministe associé à $\mathcal{C}$ pour définir ces différents niveaux, générer les motifs candidats et les élaguer. Ce travail illustre la difficulté d'intégrer certaines contraintes dans le processus de fouille.

Face aux limites des stratégies de type *Apriori*, l'algorithme *Spade* a été proposé dans [Zak01]. Il s'agit d'une adaptation de l'algorithme *Eclat* [Zak00b] à des données et des motifs séquentiels. Cet algorithme s'appuie sur une représentation des données sous forme verticale (*vertical id-list*) qui permet une génération des motifs et un calcul de leur fréquence de manière incrémentale sans avoir à accéder à toutes les données. Le tableau 2.2 présente une représentation verticale des données pour les *items* a, b et c du tableau 2.1. Cette représentation donne la position de chaque *item* dans les séquences en entrée. Dans cette approche, un motif de taille k (i.e. une séquence avec k *items*) est obtenu en faisant la jointure temporelle des listes verticales d'identifiants associées à deux motifs de taille $k-1$ partageant le même préfixe. La fréquence du motif est simplement le nombre d'identifiants différents dans la nouvelle liste verticale issue de la jointure temporelle. Par exemple, les listes verticales de $< \{a\} >$ et de $< \{c\} >$(tableau 2.2) permettent de générer les motifs $< \{a, c\} >$ et $< \{a\}, \{c\} >$ ainsi que leur liste verticale (figure 2.1). Dans ces listes verticales, les positions représentent les positions des derniers *itemsets* des séquences. De la même manière, les listes verticales de $< \{a, c\} >$ et $< \{a\}, \{c\} >$ permettent de générer le motif $< \{ac\}, \{c\} >$ ainsi que sa liste verticale. Afin de limiter l'espace mémoire nécessaire, *Spade* effectue en plus un partitionnement de l'espace de recherche basé sur le préfixe des motifs. Deux motifs sont dans la même partition s'il partage le même préfixe. Chaque partition ainsi constituée est traitée indépendamment en mémoire. Ces principes sont très généraux et peuvent être appliqués dans le cadre d'un parcours en profondeur ou en largeur de l'espace de recherche (les deux variantes ont été proposées par les auteurs). En plus de la fréquence, les auteurs ont aussi intégré à leur algorithme des contraintes temporelles (longueur, fenêtre glissante et intervalle entre deux évènements), des contraintes sur les *items* étudiés et des contraintes visant à filtrer les motifs discriminants par rapport à des classes d'intérêt [Zak00a]. Toutefois, l'ajout de certaines contraintes a un impact sur la stratégie mise en place dans l'algorithme. Par exemple, la contrainte d'intervalle maximum ne permet plus de partitionner l'espace de recherche et donc de limiter l'espace mémoire utilisé (toutes les séquences fréquentes de taille deux doivent être stockées en mémoire).

Cette représentation verticale a aussi été utilisée dans d'autres travaux tels que [AFGY02, OPS04, FVGCT14]. Par exemple, [AFGY02] proposent l'algorithme *SPAM* basé sur le même principe mais sur une structure de données sensiblement différente. Cet algorithme effectue un parcours en profondeur de l'espace de recherche à partir d'une représentation verticale des données encodée sous forme de vecteurs de bits (*vertical bitmap*). Cette structure de données a l'avantage de consommer moins de mémoire. Elle diminue aussi de manière importante le coût

a	
SID	**Temps/Positions**
1	1, 2
2	1, 2
3	1, 2
4	2
5	

b	
SID	**Temps/Positions**
1	2
2	
3	
4	3
5	1

c	
SID	**Temps/Positions**
1	1, 3
2	1, 2
3	1, 2, 3
4	1
5	2

TABLE 2.2 – Listes verticales d'identifiants (*vertical id-lists*) pour les *items* a, b, et c

<{a,c}>	
SID	**Temps/Positions**
1	1
2	1, 2
3	1, 2
4	
5	

<{a},{c}>	
SID	**Temps/Positions**
1	3
2	2
3	2, 3
4	
5	

➔

<{ac},{c}>	
SID	**Temps/Positions**
1	3
2	2
3	2, 3
4	
5	

FIGURE 2.1 – Génération du motif $< \{ac\}, \{c\} >$ à partir des listes verticales d'identifiants de $< \{a, c\} >$ et $< \{a\}, \{c\} >$

d'intersection de deux listes verticales, i.e. le comptage de la fréquence. L'algorithme *CCSM* (*Cache-based Constrained Sequence Miner*) proposé dans [OPS04] reprend aussi le principe de liste verticale de *Spade* avec un parcours par niveau de l'espace de recherche. Ils intègrent à la fois la contrainte de fréquence minimale et celle d'intervalle minimum/maximum. Face aux limites de *Spade* par rapport à la contrainte d'intervalle maximum, ils proposent d'utiliser un "cache" pour stocker des listes verticales intermédiaires, ce qui permet de limiter le nombre de jointures temporelles effectuées ainsi que la quantité mémoire consommée. Récemment, ces stratégies (*Spade* et *SPAM*) ont encore été améliorées dans [FVGCT14] grâce au concept d'élagage de co-occurrences. Cet élagage consiste à éliminer directement un motif généré si ses deux derniers *items* ne constituent pas une séquence fréquente. Pour pouvoir réaliser ce test efficacement, les séquences fréquentes de taille deux sont stockées dans une structure de données appelée $CMAP$ (*co-occurrence map*) au début de l'algorithme. Cette optimisation permet en pratique d'éliminer un grand nombre de motifs candidats sans avoir à calculer leur liste verticale d'identifiants.

De manière générale, les méthodes basées sur ce principe "générer-tester" souffrent d'un même problème : elles génèrent beaucoup de motifs non intéressants pour ensuite les élaguer. Initialement, ce problème a été mis en avant pour la fouille d'*itemsets* fréquents, ce qui a donné naissance à des stratégies différentes basées sur la notion de *pattern growth*. L'algorithme *FP-Growth* [HPY00] est le premier avoir mis en place cette stratégie pour extraire des *itemsets* fréquents. Dans ce travail, les *items* fréquents sont utilisés pour projeter récursivement la base de données et étendre les préfixes des motifs dans cet ensemble de bases de données plus petites. Pour faire les projections efficacement, l'approche s'appuie sur une structure de données arborescente appelée *fp-tree*. Ce principe a été repris dans d'autres approches et d'autres domaines de motifs (dont les séquences). *PrefixSpan* est probablement l'algorithme d'extraction de motifs séquentiels le plus connu basé sur cette stratégie [PHMA$^+$01]. Il effectue aussi un parcours en profondeur de l'espace de recherche basé sur des projections successives des données combinées avec des extensions des préfixes des motifs. L'algorithme ne considère à chaque fois que le préfixe d'un seul motif séquentiel et projette uniquement les séquences postfixes correspondantes dans les données en entrée. Le préfixe est ensuite étendu à partir des *items* fréquents locaux de la base de données

$< \{a\} >$

SID	Séquence projetée
1	$< \{-, c\}, \{a, b\}, \{c, f\}, \{e, f\} >$
2	$< \{-, c\}, \{a, c\}, \{d, f\} >$
3	$< \{-, c\}, \{a, c, e\}, \{c, e, f\}, \{f\} >$
4	$< \{-, d\}, \{b, e\}, \{e\} >$

➔ $< \{a\}, \{a\} >$

SID	Séquence projetée
1	$\{c, f\}, \{e, f\} >$
2	$< \{-, c\}, \{f\} >$
3	$< \{-, c, e\}, \{c, e, f\}, \{f\} >$

➔ $< \{a\}, \{a, c\} >$

SID	Séquence projetée
2	$< \{f\} >$
3	$< \{-, e\}, \{c, e, f\}, \{f\} >$

➔ $< \{a\}, \{a, c\}, \{f\} >$

SID	Séquence projetée
3	$< \{f\} >$

FIGURE 2.2 – Projections successives des données en fonction du préfixe d'un motif

projetée, et celle-ci est à nouveau projetée en fonction de ces extensions. Reprenons l'exemple de base de données de séquences du tableau 2.1. L'algorithme va d'abord rechercher les *items* fréquents. Si la fréquence minimale est deux, tous les items sont fréquents dans cet exemple ($a : 4$, $b : 3$, $c : 5$, $d : 3$, $e : 3$, $f : 4$). On obtient donc les préfixes de solutions suivants : $< \{a\} >$, $< \{b\} >$, $< \{c\} >$, $< \{d\} >$, $< \{e\} >$ et $< \{f\} >$. Ensuite, l'algorithme étudie le premier préfixe $< \{a\} >$ et va projeter la base de données par rapport à celui-ci. Les autres préfixes seront étudiés lorsque tous les motifs fréquents ayant pour préfixe $< \{a\} >$ auront été générés. Le premier tableau de la figure 2.2 montre le résultat obtenu par cette première projection. L'algorithme extrait les *items* fréquents dans cette base projetée et les utilise pour étendre le préfixe de la manière suivante : $< \{a\}, \{a\} >$, $< \{a, c\} >$, $< \{a\}, \{c\} >$, $< \{a\}, \{e\} >$ et $< \{a\}, \{f\} >$. L'algorithme projette ensuite la base de données par rapport à $< \{a\}, \{a\} >$ et répète les opérations récursivement. Le résultat de cette deuxième projection et des projections suivantes sont présentés dans la suite de la figure 2.2. A noter que les *items* non fréquents localement sont progressivement supprimés des bases projetées (p.ex. suppression de b et de d après la première projection). Par ailleurs, les *items* sont triés par ordre lexicographique afin de ne pas générer deux fois le même motif (cet ordre est aussi pris en compte lors des projections). De par cette stratégie, l'algorithme génère uniquement des motifs apparaissant dans les données (ce qui n'est généralement pas le cas des stratégies "générer-tester"). Toutefois, la création des bases de données projetées a un coût non négligeable (en temps et en mémoire). Afin de réduire le nombre de projections et leur taille, les auteurs introduisent un nouveau type de projection (double niveau) basée sur une matrice stockant la fréquence de toutes les séquences de taille deux. De plus, un opérateur de pseudo-projection est aussi proposé pour accélérer les traitements lorsque les données peuvent tenir en mémoire. Par cette technique, la base de donnée projetée est réduite à un ensemble de pointeurs vers les données initiales.

Beaucoup d'autres algorithmes appliquent la même stratégie tels que [SK02, LKL07, Che10]. Par exemple, [SK02] proposent l'algorithme *SLPMiner* basé sur un parcours en profondeur de l'espace de recherche et des projections successives des données. Les motifs recherchés sont toujours des sous-séquences fréquentes mais le seuil de fréquence minimale varie en fonction de la taille des motifs. D'après les auteurs, les motifs avec peu d'*items* sont intéressants s'ils ont une

fréquence élevée, alors que les longs motifs restent intéressants même si leur fréquence est relativement faible. Ils proposent donc une approche permettant d'extraire ces deux types de motifs en même temps. Toutefois, cette contrainte ne permet plus d'élaguer un motif si sa fréquence est inférieure au seuil. Les auteurs identifient donc de nouvelles propriétés basées sur la fréquence du motif, sa longueur et la longueur des séquences dans les bases projetées. Ces propriétés permettent d'élaguer des bases projetées complètes ou certaines parties (séquences et *items*), et ainsi de garantir le passage à l'échelle de l'algorithme. [LKL07] proposent une approche pour extraire des sous-séquences intéressantes à partir de traces de logiciels (des collections de séquences d'instructions). Une sous-séquence est intéressante si elle se répète. A cause des boucles, ces répétitions peuvent apparaître plusieurs fois dans une même séquence/trace. Les auteurs doivent donc étendre la définition de fréquence d'un motif à toutes les occurrences apparaissant dans les données. Afin d'avoir une contrainte (anti-)monotone pouvant être exploitée pour élaguer l'espace de recherche, seules les occurrences disjointes sont considérées. L'extraction de ces motifs est faite par une adaptation de l'algorithme *PrefixSpan* [PHMA+01] à ce contexte particulier. Les auteurs adaptent notamment les projections réalisées afin de prendre en compte les différentes occurrences dans une même séquence. [Che10] proposent une nouvelle structure de données basée sur un graphe orienté acyclique (*UpDown Directed Acyclic Graph*) pour améliorer l'efficacité des stratégies de type *pattern growth*. A chaque projection de la base de données, cette structure de données permet d'étendre la fin et le début d'un motif, alors que l'approche classique permet uniquement d'ajouter un *item* à la fin. Une séquence de taille k peut ainsi être générée en $log_2 k + 1$ projections au lieu de k.

Extraire des représentations condensées des séquences fréquentes Le nombre de solutions extrait par les algorithmes d'extraction de motifs peut être très grand (plus important que les données en entrée). Face à ce problème, des représentations condensées ont été proposées dans le cadre de la fouille d'*itemsets* puis étendues aux séquences.

L'une des premières a avoir été étudiée est l'ensemble des motifs maximaux. En effet, il est possible de connaître tous les motifs fréquents à partir des motifs fréquents maximaux (la fréquence étant monotone décroissante). Toutefois, cette représentation n'est pas sans perte d'information car il n'est pas possible de connaître leur fréquence exacte sans accéder aux données. Plusieurs algorithmes ont été proposés afin d'adapter les stratégies précédentes à l'extraction des séquences maximales. Par exemple, [LC05] utilisent la même stratégie par niveau que GSP [SA96b]. Toutefois, au lieu de compter la fréquence des motifs candidats dans toutes les données, ils utilisent un échantillon de celles-ci et recherchent des motifs non fréquents permettant d'élaguer l'espace de recherche. Les motifs générés à la fin de cette approche par niveau sont "au-dessus" des motifs fréquents maximaux. Une approche descendante (*top-down*) est donc ensuite utilisée pour extraire les maximaux. Dans [FVWT13], les auteurs adaptent la stratégie de $PrefixSpan$ [PHMA+01] pour extraire les séquences fréquentes maximales. Ils ajoutent notamment un mécanisme permettant de déterminer efficacement si un motif est maximal sans avoir à le comparer avec les motifs trouvés dans les itérations précédentes. Ce mécanisme est basé sur le principe d'extensions maximales (*maximal-backward-extension* et *maximal-forward-extension*) proposé dans [WHL07].

La représentation condensée probablement la plus utilisée est l'ensemble des fermés fréquents (*closed patterns*) [PBTL99]. Un motif fréquent est fermé s'il n'existe par de sur-motif fréquent apparaissant dans les mêmes transactions (et donc ayant la même fréquence). L'avantage de cette représentation est d'être sans perte d'information car tous les motifs fréquents, avec leur fréquence, peuvent être générés à partir des fermés sans avoir à accéder aux données. L'algorithme *CloSpan* [YHA03] est l'un des premiers à avoir été proposé pour extraire ces motifs. Il s'agit d'une

adaptation de l'algorithme *PrefixSpan* [PHMA+01] intégrant deux étapes supplémentaires afin d'extraire les séquences fermées. La première étape s'appuie sur une propriété (*early termination by equivalence*) pour éviter de générer certains motifs appartenant aux mêmes transactions qu'un motif déjà étudié et partageant le même suffixe. La deuxième étape parcourt l'ensemble des motifs extraits afin d'éliminer ceux qui ne sont pas fermés. Une limite de cet algorithme est de devoir conserver en mémoire l'ensemble des motifs fermés extraits au cours de son exécution. Face à ce problème, [WHL07] proposent l'algorithme $BIDE$ basé sur une stratégie similaire à $PrefixSpan$ [PHMA+01]. Toutefois, il extrait les motifs fermés sans avoir besoin de conserver en mémoire tout l'historique des motifs déjà testés. Il s'appuie pour cela sur deux propriétés des extensions (*bi-directional extension closure checking* et *back scan pruning*) pour savoir si un motif est fermé et élaguer les motifs déjà testés. Intuitivement, un motif n'est pas fermé s'il peut être étendu par un *item* sans pour autant changer de fréquence. De la même manière, un motif S est inclus dans un motif déjà généré s'il existe un item apparaissant "avant" S dans chacune des transactions en entrée. Ces vérifications sont effectuées à partir des *items* fréquents de la base projetée et d'informations sur l'apparition des items dans les séquences en entrée. Plus récemment, [GCMG13] proposent un algorithme (appelé $ClaSP$) plus performant que les deux précédents en s'appuyant sur une représentation verticale des données. Cet algorithme en profondeur combine la stratégie de l'algorithme $SPADE$ [Zak01] avec les propriétés mises en avant dans $CloSpan$.

Au delà des fermés, d'autres représentations condensées ont été étudiées pour les motifs séquentiels. On retrouve par exemple les générateurs (encore appelés libres) définis dans le cadre de la fouille d'*itemsets* [PBTL99, BBR03, CCL05]. Un motif fréquent est un générateur s'il n'existe par de sous-motif fréquent apparaissant dans les mêmes transactions (et donc ayant la même fréquence). Les fermés et les générateurs délimitent des classes d'équivalence de motifs (par rapport aux transactions en entrée), et sont utilisés pour générer des règles d'association exactes [PBTL99]. Les générateurs étant les plus petits motifs des classes d'équivalences, ils sont souvent préférés aux fermés pour des tâches de classification. Tout comme pour les motifs fermés, les stratégies basées sur des projections successives (*pattern growth*) et celles basées sur des représentations verticales ont été adaptées à ce contexte. Par exemple, [LKL08] adaptent l'algorithme $CloSpan$ [YHA03] et utilisent une propriété permettant d'élaguer des sur-séquences d'un générateur. Dans ce même contexte, [FVGŠH14] adaptent l'algorithme $SPAM$ [AFGY02] et les stratégies proposées dans [WHL07, FVGCT14] pour élaguer l'espace de recherche.

L'un des problèmes des précédentes contraintes est d'extraire beaucoup de motifs relativement similaires et triviaux. Des travaux se sont donc intéressés à d'autres contraintes ne dépendant pas uniquement de la fréquence. Dans le cadre de la fouille d'*itemsets*, [VVLS11] ont proposé l'algorithme $Krimp$ pour extraire les *itemsets* compressant le mieux les données. Pour trouver ces motifs, ils utilisent le principe de description de longueur minimale (*minimum description length* ou MDL) [Grü07]. Soit un ensemble de modèles $\mathcal{H}$. D'après ce principe, le meilleur modèle $H \in \mathcal{H}$ est celui qui minimise $L(H) + L(D|H)$, où $L(H)$ est la longueur en bits de H et $L(D|H)$ est la longueur en bits des données encodées avec H. L'objectif est donc de trouver des motifs minimisant la taille de l'encodage de la base de données en entrée. Les auteurs s'appuient pour cela sur la fréquence des *itemsets*. Plusieurs travaux ont mis en avant l'intérêt de ces motifs pour la classification, l'identification de composants et la détection de changements. Ce principe a été étendu aux motifs séquentiels dans [LMFC12, TV12, LMFC14]. L'une des difficultés dans ce cadre est de pouvoir prendre en compte les répétitions de motifs dans une même séquence en entrée et leurs entrelacements. [LMFC12] utilisent pour cela le même encodage basé sur un dictionnaire que dans [VVLS11]. Soit un ensemble de motifs H (un "dictionnaire"), chaque séquence de la base de données est encodée en remplaçant les occurrences de chaque motif $P \in H$ par un pointeur vers le

dictionnaire. Pour obtenir efficacement l'ensemble de motifs compressant le mieux les données, les auteurs proposent une heuristique (le problème étant NP-difficile). L'algorithme glouton proposé étend un motif solution jusqu'à ce qu'il n'améliore plus la compression des données. Il commence par un motif vide et l'étend avec l'*item* le plus fréquent. Si l'extension améliore la compression, les données sont projetées en fonction de celle-ci et une autre extension est effectuée. Sinon, le processus d'extension s'arrête. Des motifs solutions sont ainsi générés itérativement tant que la compression n'atteint pas un seuil. Les auteurs étendent leur travail dans [LMFC14]. Ils proposent un nouvel encodage basé sur le code d'Elias [WWM+99] afin de mieux gérer les intervalles entre les évènements et les entrelacements de motifs (ce que ne permet pas de faire l'encodage proposé dans [TV12]).

Extraire des motifs plus pertinents Au delà du nombre de motifs extraits, une autre problématique est la pertinence des motifs extraits pour les utilisateurs. Une grande variété de contraintes a été proposée afin de guider la fouille et d'extraire des motifs d'intérêt pour les utilisateurs. Formellement, elles se présentent souvent sous la forme d'un prédicat booléen de la forme $Q(D, P)$ avec D les données en entrée et P le motif à tester. Comme l'ont montré certains travaux présentés dans cette section, ces contraintes ont été intégrées dans les différentes stratégies, avec souvent le même enjeu : mettre en avant des propriétés de celles-ci permettant d'élaguer l'espace de recherche et de gagner en efficacité. Ces contraintes peuvent être classées en deux familles et différentes sous-catégories [MR13]. La première famille de contraintes peut être vérifiée seulement en analysant le motif. On retrouve notamment :

— les contraintes sur la longueur des motifs,

— les contraintes basées sur des modèles de motifs,

— les contraintes basées sur des opérateurs d'agrégations.

Le premier type de contraintes vise à filtrer des motifs d'une certaine longueur (inférieurs ou supérieurs). Les définitions peuvent d'ailleurs varier à ce niveau (p.ex. nombre d'*items* ou d'*itemsets*, distincts ou non). Le deuxième type de contraintes permet de spécifier quels motifs ou groupes de motifs doivent être inclus ou exclus de l'analyse, en fonction d'un modèle défini par l'utilisateur. Ce modèle peut correspondre simplement à un ensemble d'*items* prédéfinis, à des sous-séquences (ou sur-séquences) d'un motif, ou à une expression régulière. Le dernier type de contraintes est imposé sur des agrégations d'*items* (p.ex. le minimum, le maximum, ou la moyenne des valeurs associées à certains *items*).

La deuxième famille de contraintes nécessite un accès aux données en entrée et impacte généralement le calcul de la fréquence. A ce niveau, on retrouve notamment des contraintes temporelles sur les occurrences du motif telles que :

— les contraintes d'intervalle de temps entre les évènements,

— les contraintes de durées.

Les contraintes d'intervalles permettent de filtrer les occurrences d'un motif dont les évènements (ou les transactions) sont trop proches ou trop éloignées. Pour cela, des informations complémentaires sur les données en entrée peuvent être nécessaires, comme l'estampille temporelle de chaque évènement. Ces bases de séquences intégrant ces informations temporelles sont aussi appelées base de séquences temporelles. Les contraintes de durées peuvent d'ailleurs uniquement être mises en place dans ce type de base de données. Elles permettent de définir un durée minimale ou maximale entre le premier et le dernier évènement d'une occurence de motif (ou entre la première et la dernière transaction supportant le motif).

De manière similaire, d'autres types d'information ont été intégrés à la base de données de séquences tels que le poids, le prix, la quantité, ou la probabilité d'apparition associés à chaque *item* [YL06, MR11, YZC12]. Ces informations ne filtrent pas uniquement des occurrences de motifs comme précédemment. Elles modifient la définition même du prédicat déterminant l'intérêt d'un motif. Les travaux étudiant l'extraction de motifs à haute utilité (*high-utility patterns*) et ceux étudiant les bases de séquences incertaines sont notamment dans ce cas. Par exemple, [YZC12] étudient une collection de séquences, où chaque *item* est associé à un poids (ou une qualité) global et à une quantité dans chaque séquence en entrée. Les motifs recherchés sont des sous-séquences ayant une utilité élevée. La fonction indiquant l'utilité d'un motif dépend du poids des items et de leur quantité dans les données en entrée. Dans leur article, les auteurs utilisent comme exemple une fonction faisant la somme des produits poids-quantité. Un motif est solution si son utilité maximale est supérieure à un seuil. Cette contrainte n'est pas (anti-)monotone. Il n'est donc pas possible de l'utiliser directement pour élaguer l'espace de recherche. Les auteurs mettent donc en avant deux propriétés permettant de filtrer des motifs en fonction de bornes supérieures sur l'utilité des extensions possibles. Puis, ils adaptent l'algorithme $SPAM$ [AFGY02] afin d'intégrer ces données et ces contraintes. [MR11] intègrent quant à eux l'information sur le bruit et l'incertitude. Ils considèrent en entrée des bases de séquences probabilistes. Chaque séquence en entrée correspond à une source de données où les *itemsets* sont associés à des probabilités d'apparition dans la source en question. La notion de fréquence "attendue" est alors utilisée pour filtrer des sous-séquences intéressantes. Un algorithme incrémental de programmation dynamique est proposé pour calculer efficacement cette fréquence basée sur la probabilité qu'une sous-séquence soit associée à une source. Cet algorithme est à son tour intégré dans deux stratégies d'exploration de l'espace de recherche (une en largeur et une en profondeur) dérivées de *GSP* [SA96b] et de *SPAM* [AFGY02]. Les auteurs mettent aussi en avant une propriété permettant d'élaguer certains motifs candidats grâce à une borne supérieure sur la probabilité d'apparition d'une sous-séquence dans une source.

Les approches précédentes recherchent des motifs se répétant exactement de la même manière. Toutefois, comme indiqué dans [KPWD03], cette contrainte peut être trop forte dans des bases de données avec de longues séquences et du bruit. Pour cette raison, [KPWD03] recherchent des motifs fréquents approximatifs, i.e. des motifs approximativement partagés par beaucoup de séquences en entrée. De plus, ils se limitent aux motifs longs couvrant un grand nombre de motifs de petite taille. Ces motifs sont appelés motifs séquentiels de consensus (*concensus sequential patterns*). Les auteurs proposent une approche de *clustering* basée sur de multiples alignements pour extraire ces motifs. Les séquences en entrée sont tout d'abord regroupées en fonction de leur similarité (dérivée de la distance d'édition hiérarchique). Puis, l'approche construit pour chaque *cluster* une séquence pondérée enregistrant des statistiques sur l'alignement des séquences dans celui-ci. Pour finir, cette séquence est utilisée pour générer le plus long motif de consensus. Ce dernier travail met en avant une autre famille de problèmes non étudiés dans cet état de l'art : l'analyse de séries temporelles (*time series*). Dans ce cadre, l'objectif est sensiblement différent. Il ne s'agit pas d'extraire des modèles "locaux" représentant une partie des données, mais d'extraire des modèles "globaux" décrivant globalement l'ensemble. De plus, ces séries temporelles sont généralement composées de valeurs numériques. Cette famille de problème est souvent associée à des méthodes de *clustering* ou de classification supervisée, et à la définition de mesures de similarité.

Le tableau 2.3 résume les principaux travaux décrits précédemment. La colonne "Article" donne les références bibliographiques des articles et éventuellement le nom des algorithmes associés (entre parenthèses). La colonne "Motifs extraits" précise le domaine de motifs extraits. La colonne "Contraintes et Mesures" indique les contraintes et mesures utilisées pour filtrer

les motifs intéressants. La colonne "Méthode d'extraction" présente les principaux éléments de l'algorithme utilisé pour l'extraction. La colonne "Application" indique les données et applications utilisées dans les expérimentations. Les données "synthétiques" correspondent à des jeux de données générés selon des paramètres contrôlés. Les données "benchmark" sont des jeux de données publiques utilisés par la communauté pour comparer les méthodes (p.ex. données de la compétition *KDD cup*). Ces deux types de données sont généralement utilisés pour des tests quantitatifs. Les autres types de données correspondent à de réelles applications avec des études qualitatives des résultats.

TABLE 2.3 – Extraction de motifs dans des séquences : synthèse

Article	**Motifs extraits**	**Contraintes et Mesures**	**Méthode d'extraction**	**Applications**
[AS95]	séquences d'*itemsets*	fréquence min.	stratégie d'*Apriori* [AS+94]	données synthétiques
[SA96b] (*GSP*)	séquences d'*itemsets*	fréquence min., fenêtre temporelle, intervalle max., taxonomie	stratégie d'*Apriori* [AS+94], arbre de hachage	données synthétiques, données benchmark
[MCP98] (*PSP*)	séquences d'*itemsets*	fréquence min.	stratégie d'*Apriori* [AS+94], arbre préfixe	données synthétiques
[GRS99] (*SPIRIT*)	séquences d'*itemsets*	fréquence min., expression régulière	stratégie d'*Apriori* [AS+94], relaxation de contraintes, automate fini déterministe	usages du Web, données synthétiques, données benchmark
[Zak01] (*Spade*)	séquences d'*itemsets*	fréquence min., longueur max., fenêtre glissante, intervalle max.	parcours en profondeur/largeur, représentation verticale [Zak00b], partitionnement de l'espace de recherche	données synthétiques, plans d'évacuation d'une ville
[AFGY02] (*SPAM*)	séquences d'*itemsets*	fréquence min.	parcours en profondeur, vecteur de bits vertical	données synthétiques
[OPS04]	séquences d'*itemsets*	fréquence min., intervalle min./max.	parcours par niveau, représentation verticale, cache	données synthétiques
[FVGCT14]	séquences d'*itemsets*	fréquence min.	stratégie de *Spade* [Zak01] ou stratégie de *SPAM* [AFGY02], *co-occurrence map*	données benchmark
[PHMA+01] (*PrefixSpan*)	séquences d'*itemsets*	fréquence min.	parcours en profondeur, projections des données et *prefix growth* [HPY00], projection double niveaux pseudo projections	données synthétiques
[SK02]	séquences d'*itemsets*	fréquence min. variant avec la taille des motifs	stratégie de *PrefixSpan* [PHMA+01]	données synthétiques
[LKL07]	séquences d'*itemsets*	fréquence min. (nombre d'occurences disjointes)	stratégie de *PrefixSpan* [PHMA+01]	traces logicielles, données benchmark
[Che10]	séquences d'*itemsets*	fréquence min.	stratégie de *PrefixSpan* [PHMA+01], *UpDown Directed Acyclic Graph*	données benchmark
[LC05]	séquences d'*itemsets*	fréquence min., maximaux	stratégie de *GSP* [SA96b], échantillonnage et approche descendante	données synthétiques
[FVWT13]	séquences d'*itemsets*	fréquence min., maximaux	stratégie de *PrefixSpan* [PHMA+01], *bi-directional extension checking* [WHL07]	données benchmark
[YHA03] (*CloSpan*)	séquences d'*itemsets*	fréquence min., fermeture	stratégie de *PrefixSpan* [PHMA+01], *early termination by equivalence*	données synthétiques, données benchmark

Suite sur la page suivante

TABLE 2.3 – *Suite de la page précédente*

Article	Motifs extraits	Contraintes et Mesures	Méthodes d'extraction	Applications
[WHL07] (*BIDE*)	séquences d'*itemsets*	fréquence min., fermeture	stratégie de *PrefixSpan* [PHMA+01], *bi-directional extension checking back scan pruning*	données synthétiques, données benchmark
[GCMG13]	séquences d'*itemsets*	fréquence min., fermeture	stratégie de *Spade* [Zak01] combinée avec les propriétés de *CloSpan* [YHA03]	données synthétiques
[LKL08]	séquences d'*itemsets*	fréquence min., générateur	stratégie de *CloSpan* [YHA03]	données benchmarks
[FVGŠH14]	séquences d'*itemsets*	fréquence min., générateur	stratégie de *SPAM* [AFGY02], *co-occurrence map* [FVGCT14] *bi-directional extension checking* [WHL07]	données benchmarks
[LMFC12]	séquences d'*itemsets*	compression	algorithme glouton, description de longueur minimale, encodage de *Krimp* [VVLS11]	données benchmark
[LMFC14]	séquences d'*itemsets*	compression	algorithme glouton, description de longueur minimale, encodage d'Elias [WWM+99]	données synthétiques, données benchmark, analyse de mots clés de publications
[YZC12]	séquences d'*itemsets*	utilité max.	stratégie de *SPAM* [AFGY02], bornes supérieures	données synthétiques, données benchmark
[MR11]	séquences d'*itemsets*	fréquence min. "attendue"	stratégie de *GSP* [SA96b] et stratégie de *SPAM* [AFGY02], calcul incrémental de la fréquence	
[KPWD03]	séquences d'*itemsets* approximatives	fréquence min.	*clustering*, distance d'édition hiérarchique, multiples alignements	données synthétiques, données de services sociaux

2.1.2 Les graphes étiquetés, attribués et dynamiques

Dans un grand nombre d'applications, les données peuvent être représentées sous la forme de graphes. Ils sont par exemple utilisés pour représenter la structure de molécules en médecine ou l'organisation des réseaux sociaux. Dans ce contexte, l'extraction de motifs dans des graphes a été au centre de beaucoup de travaux ces dernières années. Les stratégies développées sont souvent similaires à celles proposées pour extraire les *itemsets* ou les *séquences* fréquentes. Un certain nombre d'approches s'appuient même sur une transformation des graphes sous forme de séquences (basé sur le parcours en profondeur du graphe). La construction des motifs est aussi souvent similaire : ajout d'un *item* dans l'*itemset* d'un noeud existant ou extension du graphe par un nouveau noeud associé à un *item*. Le passage à l'échelle est néanmoins beaucoup plus difficile en raison de la complexité structurelle des graphes. Pour cette raison, un grand nombre d'hypothèses, de contraintes, et de domaines de motifs différents ont été étudiés. La suite de cette section présente les principales approches pour fouiller ces graphes, en distinguant notamment les différentes données en entrées, les différents domaines de motifs étudiés et les différentes contraintes considérées.

Extraire des motifs dans une collection de graphes étiquetés Les premiers travaux dans ce domaine ont considéré en entrée une collection de graphes étiquetés et ont essayé d'en extraire des sous-graphes étiquetés fréquents. La figure 2.3 présente un exemple de graphe où chaque noeud est étiqueté par une valeur (a, b, c, e, ou f). Dans ce cadre, la fréquence d'un sous-graphe est généralement définie comme étant le nombre de graphes en entrée qui le contiennent.

Ainsi, ces approches limitent le coût du test d'isomorphisme (un problème NP-complet) en ne recherchant qu'une seule occurence du sous-graphe dans chaque graphe en entrée.

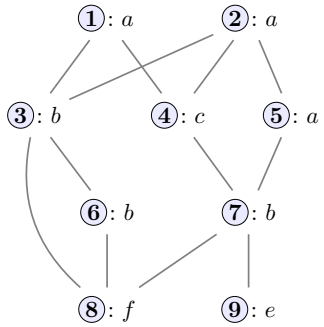


FIGURE 2.3 – Exemple de graphe étiqueté

[IWM00] proposent l'algorithme *AGM* pour extraire des sous-graphes fréquents (et connectés) dans une collection de graphes étiquetés (noeuds et/ou arêtes). Dans ce travail, les auteurs se focalisent sur les motifs représentant des sous-graphes induits (*induced subgraph*), i.e. des sous-graphes respectant la relation parent-enfant. Les motifs fréquents sont extraits suivant un parcours par niveau de l'espace de recherche dérivé d'*Apriori* [AS+94]. L'algorithme s'appuie sur une transformation des graphes sous forme de séquences (*canonical code*) basée sur la matrice d'adjacence. Cette représentation a certaines propriétés facilitant la génération et l'indexation des motifs, tout en limitant le nombre de motifs testés. Les motifs sont ainsi générés de manière unique en ajoutant un noeud à la fois. Au final, l'algorithme est utilisé pour caractériser des substances cancérigènes dans une base de données de composants chimiques. Par la suite, [KK01] proposent une nouvelle structure de données et différentes optimisations pour améliorer l'efficacité de cet algorithme (notamment la génération des motifs candidats et le comptage de la fréquence).

Une stratégie par niveau de type générer-tester, telle que celle présentée précédemment, génère un grand nombre de motifs candidats et multiplie les accès à la base de données pour compter la fréquence. Son efficacité est donc limitée lorsqu'il s'agit d'extraire de "longs" motifs. Face à ce problème, [YH02] proposent un nouvel algorithme, appelé *gSpan*, pour extraire des sous-graphes fréquents induits (*induced*) et inclus (*embedded*). Ces derniers sont des sous-graphes respectant la relation ancêtre-descendant. La stratégie de l'algorithme est relativement similaire à celle utilisée dans l'algorithme *PrefixSpan* [PHMA+01] pour la fouille de séquences (*pattern growth strategy*). *gSpan* effectue un parcours en profondeur de l'espace de recherche associé à des projections successives des données en entrée. L'intérêt de cette approche est de limiter la quantité de mémoire consommée à un temps donné et d'accélérer le comptage de la fréquence. Les graphes sont transformés en code unique (une séquence) basé sur leur parcours en profondeur dans l'ordre lexicographique (*minimum DFS codes*), puis étendu récursivement. L'approche combine ainsi test d'isomorphisme et extension des motifs, ce qui la rend plus efficace. Comme nous le verrons dans la suite, cet algorithme est utilisé ou adapté dans un grand nombre de travaux. Ces différentes variantes diffèrent notamment dans les contraintes intégrées en plus de la fréquence minimale. Par exemple, [JCZ10] l'utilisent pour extraire des sous-graphes dont la fréquence est pondérée en fonction des arêtes qui les composent. Les auteurs introduisent trois contraintes liées au poids total du motif, à son affinité et à son utilité. L'affinité pondère la fréquence du motif par rapport au nombre de noeuds et considère l'homogénéité des poids. L'utilité pondère la fréquence par un nombre inversement proportionnel à la similarité entre les noeuds du motif

(d'après une distance de Jaccard) et considère son poids par rapport au poids de tous les graphes dans lesquels il apparaît.

Plusieurs autres algorithmes ont été proposés pour extraire plus efficacement des sous-graphes tels que [NK04, WHLS06, DLLS15]. [NK04] partent du constat que le test d'isomorphisme peut être effectué en temps polynomial pour certaines sous-familles de graphes tels que les chemins et les arbres libres (*free trees*). Ils proposent donc un algorithme en profondeur, appelé *GASTON*, qui commence par rechercher les chemins fréquents, puis ces chemins sont utilisés pour générer des arbres fréquents, et pour finir ces arbres sont utilisés pour générer des graphes fréquents. Grâce au parcours en profondeur, l'algorithme peut stocker toutes les occurrences (*embeddings*) d'un motif donné et utiliser cette information pour accélérer le calcul de la fréquence lors des extensions. [WHLS06] proposent une approche basée sur un partitionnement de l'espace de recherche pour extraire des sous-graphes fréquents dans une collection de graphes étiquetés. Classiquement, les algorithmes dans ce domaine font l'hypothèse que la base de données peut être chargée en mémoire, ce qui facilite et accélère le comptage de la fréquence. Dans ce travail, les auteurs considèrent au contraire que dans beaucoup de cas, la base de donnée en entrée ne peut pas tenir en mémoire. Ils proposent donc de diviser les données en entrée en plusieurs partitions de taille plus petites, et d'extraire les sous-graphes fréquents dans ces partitions en utilisant l'algorithme *GASTON* [NK04]. Les résultats de ces extractions sont ensuite combinés afin de générer l'ensemble complet des solutions. Récemment, [DLLS15] présentent un nouveau cadre pour extraire des sous-graphes fréquents induits. Les auteurs se concentrent plus particulièrement sur le problème du test d'isomorphisme de sous-graphes. Ils proposent d'introduire un biais dans l'opérateur de projection et d'utiliser un opérateur polynomial permettant une interprétation sémantiquement valide de la structure. Leur approche est basée sur une technique de propagation de contraintes (*local consistency*) développée dans le domaine de la programmation par contraintes. Cette technique a été intégrée dans un algorithme par niveau de type *Apriori* [AS$^+$94] et dans un algorithme effectuant un parcours en profondeur avec des extensions successives de type *PrefixSpan* [PHMA$^+$01]. De nombreux algorithmes ont été proposés pour extraire des sous-graphes fréquents dans une collection de graphes étiquetés. Des descriptions de ces algorithmes peuvent être trouvées dans les revues effectuées dans [WM03, JCZ13].

Par ailleurs, beaucoup de contributions se focalisent uniquement sur une classe spécifique de graphes (p.ex. les arbres, les graphes acycliques, ou les graphes orientés). Comme nous l'avons vu précédemment, l'idée est de tirer partie de leurs spécificités (propriétés théoriques) pour développer des algorithmes optimisés pour ces représentations. Par exemple, des travaux se sont intéressés aux graphes orientés acycliques (*Directed Acyclic Graphs* ou DAG) et à leur multiples applications [WDW$^+$08, NNP$^+$09, GS10, MSSR10, FAL$^+$13]. Par exemple, [WDW$^+$08] étudient la fouille de DAG pour optimiser la compression de codes en programmation logicielle. Ils proposent une approche pour extraire des motifs fréquents dans une collection de DAG étiquetés. Les motifs sont des sous-DAG induits, potentiellement non connexes et avec des racines multiples. Ils soulignent que la fouille de DAG est aussi un problème NP-complet car le test d'isomorphisme de sous-DAG est aussi NP-complet. Face à cette limite, ils proposent d'encoder les DAG sous une forme canonique pour éviter des tests superflus. Ce code correspond à une séquence de noeuds du sous-DAG basé sur le niveau topologique de chaque noeud. Cette représentation contient toutes les informations du sous-DAG, mais d'une façon facilitant l'énumération et le test des motifs. Les auteurs intègrent cette approche dans un algorithme par niveau, appelé *DAGMA*, qui est plus rapide et utilise moins de mémoire que l'algorithme général *gSpan* [YH02]. Dans un contexte différent, [FAL$^+$13] recherchent des motifs d'interactions entre personnes durant des réunions. Leur base de données est constituée d'un ensemble de flux d'interactions entre personnes, chacun étant représenté sous la forme d'un DAG étiqueté. Dans ce travail, les noeuds ont des poids pour

distinguer le niveau hiérarchique des participants. Leur problème est donc de trouver les sous-DAG fréquents pondérés dans une collection de DAG étiquetés et pondérés. Face à ce problème, ils adaptent le travail de [WDW+08] pour intégrer les poids dans la mesure d'intérêt et lors de l'extraction.

Extraire des motifs dans un unique graphe étiqueté Les travaux présentés précédemment étudient l'extraction de motifs fréquents dans une collection de graphes (*graph-transaction setting*). Toutefois, dans de nombreux domaines d'application, il y a uniquement un seul (grand) graphe en entrée (*single-graph setting*). Ces deux cadres partagent des propriétés mais les algorithmes développés pour analyser une collection de graphes ne peuvent pas nécessairement être utilisés pour analyser un seul graphe, alors que l'inverse est vrai [KK05]. De plus, l'un des problèmes lors de la fouille d'un seul graphe est le test d'isomorphisme. Dans le cadre d'une base de données composée d'une collection de graphes, la fréquence d'un motif est généralement définie comme le nombre de graphes en entrée (de "transactions") contenant le motif. Le test d'isomorphisme est donc arrêté dès qu'une occurrence est trouvée dans la transaction. Dans le cadre d'une base de données composée d'un graphe unique, il est nécessaire de trouver toutes les occurrences du motif, ce qui a un impact important sur les performances. Par ailleurs, il est plus difficile de définir la mesure de fréquence dans ce cadre du graphe unique. Plusieurs travaux ont étudié cette problématique [KK05, FB07, BN08, JXWT09]. Ils définissent généralement une mesure de fréquence basée sur le nombre d'occurrences mais cela n'est pas simple : plusieurs occurrences peuvent s'entrelacer, aboutissant à une mesure de fréquence non monotone. Or, il s'agit de la principale propriété utilisée par les algorithmes d'extraction pour élaguer l'espace de recherche et passer à l'échelle.

Plusieurs algorithmes ont été proposés pour extraire des motifs dans un unique graphe étiqueté [CH94, MHMW00, VGS02, KK05, GSV06, LZY10, MSSR10]. Dans [CH94], les auteurs cherchent des sous-structures communes permettant de compresser des données structurées en entrée. Trois applications sont considérées : l'analyse d'un composant chimique, l'analyse d'un circuit électronique et l'analyse d'une scène (image). La base de données est représentée sous la forme d'un unique graphe étiqueté (orienté ou non), où les objets de la base sont les noeuds et les relations entre ceux-ci sont les arêtes. Leur objectif est de trouver les sous-structures, i.e. les sous-graphes, compressant le mieux les données en entrée. L'évaluation de chaque motif est faite à partir du principe de description de longueur minimale (*minimum description length*) et d'autres contraintes données par les experts du domaine. Un algorithme glouton, appelé *SUBDUE*, a été proposé. Il s'appuie sur une recherche en faisceau sous contrainte (*constrained beam search*) pour trouver les solutions. [MHMW00] développent un autre algorithme glouton similaire pour cette problématique. [KK05] proposent deux algorithmes pour extraire des sous-graphes fréquents inclus (*embedded*) dans un unique grand graphe épars. Le premier effectue un parcours en largeur de l'espace de recherche alors que le deuxième effectue un parcours en profondeur. Tout comme les travaux précédents, ils utilisent un code séquentiel (appelé *canonic label*) pour générer et tester les motifs plus efficacement. Une particularité de leur approche est de rechercher des sous-graphes fréquents pour lesquels les inclusions ont des arêtes disjointes (problème de la recherche d'un stable de taille maximum). Cette définition de fréquence est choisie car elle préserve la propriété de monotonie de la contrainte. [GSV06] introduit un nouvel algorithme par niveau s'appuyant sur des chemins avec des arêtes disjointes. La principale originalité de leur approche est de s'appuyer sur ces chemins pour étendre les motifs, au lieu de les entendre par un noeud ou une arête à la fois. Le nombre d'itérations et de motifs générés est ainsi diminué, ce qui augmente l'efficacité. Dans [LZY10], toutes les occurrences d'un sous-graphe sont considérées,

même celles qui se chevauchent. La mesure classique de fréquence n'est donc plus monotone (décroissante). Face à ce problème, les auteurs proposent une autre contrainte liée à la fréquence : un motif est fréquent si et seulement si les fréquences minimale et maximale de ses noeuds sont supérieures à un seuil défini par l'utilisateur. Ils proposent un algorithme, appelé *DESSIN*, pour extraire ces motifs. Il est basé sur un parcours en profondeur de l'espace de recherche ainsi que sur une structure de données permettant d'indéxer et de rechercher efficacement les occurrences d'un motif.

Extraire des représentations condensées des sous-graphes étiquetés fréquents Tout comme pour la fouille de motifs séquentiels, la notion de fermeture a aussi été étudiée dans le cadre d'une collection de graphes. L'objectif reste le même : réduire le nombre de solutions sans pour autant perdre d'information. Plusieurs algorithmes ont été développés pour extraire des motifs fermés fréquents dans une collection de graphes étiquetés tels que [YH03, TTN+07, BHPG11, OE11]. [YH03] proposent un algorithme pour extraire des sous-graphes fermés fréquents. Ils soulignent que les optimisations et les propriétés développées pour les *itemsets* et les séquences ne sont pas valables pour les graphes. Face à ce problème, ils introduisent de nouveaux concepts (*equivalent occurrence* et *early termination*) pour améliorer l'élagage des motifs. Ces propriétés ont été intégrées dans l'algorithme *CloseGraph*, basé sur l'algorithme *gSpan* [YH02] et sa stratégie d'extensions en profondeur. [BHPG11] étudient l'extraction de graphes fermés dans des flux. La donnée en entrée est une séquence de graphes, chacun étant associé à un poids. Ils définissent la fréquence d'un sous-graphe comme la somme des poids des graphes en entrée qui le contiennent. Trois algorithmes incrémentaux basés sur un parcours en profondeur sont proposés pour extraire ces motifs. Ils diffèrent principalement dans leur façon de gérer les nouvelles données (mise à jour des motifs suivant des lots de nouvelles données, des fenêtres glissantes ou en continu).

L'extraction de sous-DAG fermés fréquents inclus (*embedded*) dans une collection de DAG étiqueté a aussi été étudiée dans [TTN+07]. Dans ce travail, chaque DAG doit avoir des étiquettes distinctes. Même avec cette restriction, les auteurs montrent que l'extraction de tels motifs dans un réseau de gènes composé de 100 graphes (réseaux) et de 50 étiquettes (gènes) n'est pas trivial. Un algorithme en deux étapes, appelé *DigDag*, a été développé pour trouver ces motifs. La première étape extrait les arêtes fermées fréquentes respectant la relation ancêtre-descendant en utilisant l'algorithme *LCM* [UAUA04] initialement développé pour les *itemsets*. La seconde étape combine ces arêtes pour découvrir les sous-DAG fermés fréquents.

L'extraction d'une représentation condensée sans perte d'information a été peu étudiée dans un unique graphe. Seuls les motifs fréquents maximaux ont été considérés dans [FGCOMT14]. Le problème est plus difficile. Par exemple, la définition classique de fermeture, et ses propriétés, ne sont pas applicables directement dans le cadre du graphe unique, car elles sont fortement liées au concept de transaction.

Extraire des motifs dans des graphes attribués Dans beaucoup d'applications, les graphes étiquetés ne suffisent pas à capturer toutes les informations disponibles. En particulier, associer un noeud avec une seule étiquette (attribut ou *item*) est une importante limitation. Les graphes attribués ont été introduits pour résoudre ce problème. Un graphe attribué est un graphe où chaque noeud peut être étiqueté par plusieurs attributs (un *itemset*). La figure 2.4 présente un exemple de graphe attribué où chaque noeud est étiqueté par un ensemble de valeurs. Toutefois, fouiller de tels graphes entraîne une explosion combinatoire (en raison de l'exploration conjointe des espaces de recherche des graphes et des *itemsets*), ce qui soulève de nouveaux défis.

Un certain nombre de travaux sur la fouille de graphes attribués se focalisent sur la découverte

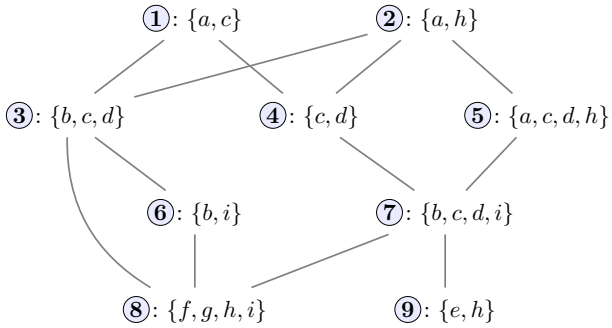


FIGURE 2.4 – Exemple de graphe attribué

de communautés (e.g. [MCRE09, FSKS10, KYW10, SMJZ12]). Ce problème est central dans plusieurs domaines tels que l'analyse de réseaux sociaux ou l'étude de systèmes biologiques. Dans les réseaux sociaux, une tâche importante est l'identification de groupes de personnes ayant de fortes interactions sociales et des intérêts similaires. Dans les systèmes biologiques, une application est l'identification de groupes de gènes avec la même expression génétique. [MCRE09] introduisent le problème de l'extraction de motifs cohésifs (*cohesive patterns*), i.e. de sous-graphes partageant les mêmes *itemsets*. Ces motifs combinent le principe de sous-graphes denses et de *subspace clusters*. Ils proposent l'algorithme *CoPaM* pour extraire ces motifs par une approche par niveau. [FSKS10] étudient des motifs similaires mais sans considérer de contrainte de densité. Leur algorithme combine une exploration en profondeur de l'espace de recherche et une structure de données de type arbre préfixe afin de trouver efficacement les motifs maximaux. Dans leurs expérimentations, ils analysent des réseaux biologiques (les voies métaboliques) pour la recherche de médicaments et analysent les collaborations dans des réseaux de citations.

Relativement peu de travaux ont considéré l'extraction de motifs fréquents dans des graphes attribués et la majeure partie d'entre eux se focalisent sur l'analyse d'une collection de graphes en entrée (*graph-transaction setting*) [BB02, BMB04, CKK04, MOO09, PFSSF17]. Dans [BB02], les auteurs analysent des sous-structures moléculaires afin de développer de nouveaux médicaments. Ils considèrent en entrée une collection de molécules et cherchent les sous-structures fréquentes permettant de discriminer des classes d'activité. Chaque molécule est modélisée par un graphe attribué. Dans ce cadre, un noeud représente un atome étiqueté par son élément chimique, sa charge et les noyaux aromatiques associés. Une arête représente une liaison entre deux atomes étiquetée par un type de liaison chimique. Les motifs sont des sous-graphes constitués d'une composante connexe. Pour extraire ces motifs, ils proposent un algorithme en profondeur, appelé *MoFa*, similaire à l'algorithme *Eclat* développé pour la fouille d'*itemsets* [ZPOL97]. Dans [BMB04], les auteurs étendent leur algorithme afin qu'il puisse extraire des sous-structures fermées fréquentes. Cet algorithme est aussi dérivé de l'algorithme *Eclat*. Tout comme [YH03], ils évitent des recherches redondantes en introduisant de nouvelles propriétés (*equivalent sibling pruning* et *perfect extension*).

[CKK04, MOO09, PFSSF17] ont étudié l'extraction de motifs fréquents dans des DAG attribués. [CKK04] analysent une base de données composée d'une collection des DAG attribués. Toutefois, ils se focalisent sur la recherche de motifs pyramidaux induits (*induced pyramid patterns*), i.e. des DAG avec un seul noeud source et respectant la relation parent-enfant. Ils proposent pour cela un algorithme similaire à *Apriori* [AS$^+$94]. [MOO09] étudient l'extraction de sous-DAG fréquent dans un unique graphe attribué. Dans ce travail, les noeuds du graphe sont associés à la fois à une étiquette et à un *itemset* quantitatif [WMM05] (i.e. un ensemble de couples

attribut-intervalle de valeurs). En conservant des étiquettes, l'extraction des motifs est simplifiée et décomposée en deux étapes : fouiller le graphe étiqueté par l'algorithme *gSpan* [YH02] et extraire les *itemsets* quantitatifs par l'algorithme *QFIMiner* [WMM05]. La mesure de fréquence utilisée est basée sur le nombre minimum d'occurrences ayant des arêtes disjointes. [PFSSF17] proposent un cadre pour extraire des sous-graphes fréquents dans des DAG enracinés. Un DAG est enraciné s'il est possible de trouver un noeud v tel qu'il existe un chemin entre v et n'importe quel autre noeud du graphe. Ce cadre peut prendre en entrée une collection de DAG ou un unique DAG. L'extraction des solutions s'appuie sur l'arbre couvrant des graphes en entrée et sur une nouvelle forme canonique des motifs. Les auteurs adaptent plus particulièrement l'algorithme proposé dans [PSFSF16] pour extraire des sous-arbres attribués. Il est basé sur un parcours en profondeur de l'espace de recherche et sur des extensions successives du chemin le plus à droite (*rightmost path extension*).

Extraction de motifs dans des graphes dynamiques La dimension temporelle a aussi été intégrée dans de nombreux travaux étudiant les graphes, notamment pour analyser les évolutions. On obtient alors des graphes dynamiques, i.e. des graphes dont les noeuds, les arêtes et les informations associées peuvent évoluer dans le temps. La figure 2.5 présente un exemple de graphe dynamique étiqueté. Elle illustre l'évolution d'un graphe (noeuds, arêtes et étiquettes) sur trois temps (t_1, t_2, et t_3). Ces travaux ont été conduits dans le cadre du graphe unique (*single-graph setting*) et dans le cadre d'une collection de graphes (*graph-transaction setting*). Toutefois, ils se limitent principalement à l'analyse des graphes dynamiques étiquetés, peu prennent en compte les graphes attribués. Par ailleurs, face à cette tâche particulièrement complexe, différentes contraintes ont été considérées pour améliorer le passage à l'échelle et la pertinence des motifs.

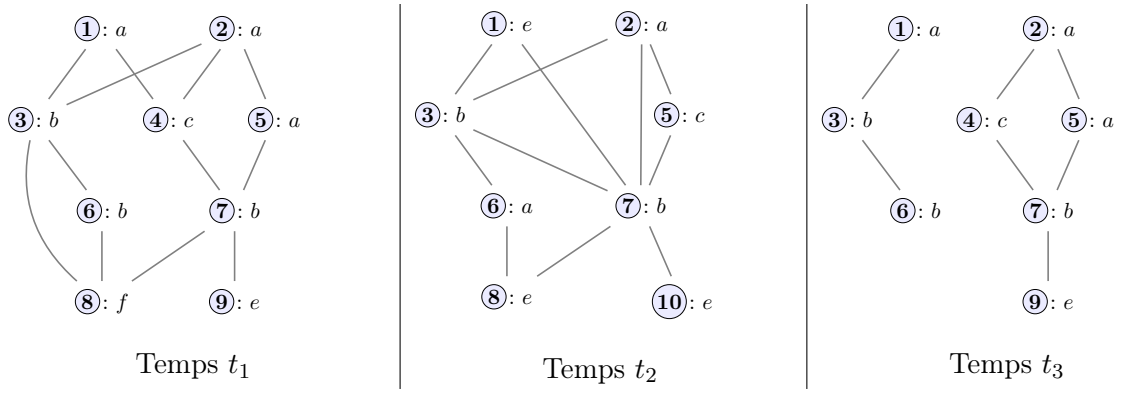


FIGURE 2.5 – Exemple de graphe dynamique étiqueté

[IW08] recherchent des transformations fréquentes dans une collection de séquences de graphes étiquetés. Pour cela, ils font l'hypothèse que seule une petite partie du graphe change entre deux temps consécutifs. Une séquence de graphes peut ainsi être représentée de manière plus compacte par la succession de transformations (insertion, suppression, etc) qui lui sont appliquées, i.e. une séquence de transformations. Ils combinent ensuite deux algorithmes existants (*AGM* [IWM00] et *PrefixSpan* [PHMA+01]) pour générer les transformations fréquentes. Pour chaque séquence en entrée, le premier génère tous les sous-graphes fréquents connexes à partir de l'union des graphes. Le deuxième recherche ensuite les sous-séquences fréquentes de transformations pour chacun de ces sous-graphes extraits. Dans [IW10], les mêmes auteurs proposent

un autre algorithme pour fouiller une collection de séquences de graphes. Contrairement au travail précédent, ils extraient des sous-séquences fréquentes de sous-graphes induits globalement connexes (i.e. seule l'union des graphes est connexe). Les motifs sont donc à la fois plus généraux et plus précis, car ils ne se limitent pas à des changement globaux sur les graphes en entrée. Pour l'extraction, l'algorithme proposé est similaire à celui proposé dans le travail précédent. Face à ce même problème, [OO09] proposent un algorithme par niveau pour extraire des séquences de sous-graphes fréquentes dans un unique graphe dynamique étiqueté. Les sous-graphes composant chaque motif doivent avoir une taille minimale. Ils doivent aussi respecter deux à deux un seuil minimum de corrélation (d'après la mesure proposée dans [TKS02] pour l'extraction de règles d'association). Afin d'assouplir cette contrainte, un nombre maximum d'exceptions dans les dépendances entre sous-graphes est également autorisé. L'efficacité de l'algorithme est améliorée grâce à une structure de données de type arborescente qui permet d'exploiter efficacement la relation de spécialisation-généralisation entre les motifs.

Beaucoup d'autres travaux ont étudié la fouille d'un unique graphe dynamique. L'approche est souvent la même : intégrer différentes contraintes pour garantir le passage à l'échelle et avoir des solutions plus pertinentes. Par exemple, [LBW08] se focalisent sur l'extraction de sous-graphes fréquents périodiques. Dans ce travail, les noeuds du graphe en entrée n'ont pas d'étiquette. La fréquence correspond au nombre de temps où le sous-graphe apparaît et la périodicité est l'écart entre deux apparitions successives (cet écart est constant). En plus de ces contraintes, les motifs doivent être fermés. Contrairement au problème classiquement étudié, celui-ci est polynomial (en temps et en espace). L'algorithme développé pour extraire les solutions s'appuie sur un parcours en largeur de l'espace de recherche basé sur une table de hachage des motifs. Suite à ce travail, d'autres approches ont été proposées afin d'extraire ces motifs plus efficacement en partitionnant l'espace de recherche et en exploitant des structures de données optimisées [ABP11, HSL17]. [BMS11] considèrent quant à eux un problème différent. Ils analysent un graphe dynamique dont les arêtes sont associées à des valeurs numériques évoluant dans le temps (un problème *NP-difficile*). Seule les valeurs associées aux arêtes changent (les noeuds et les arêtes ne changent pas). L'objectif est d'extraire les sous-graphes (connexes) maximisant le score total des arêtes entre deux temps t_i et t_j. Pour cela, ils proposent une heuristique basée sur un élagage progressif de groupes de sous-intervalles de temps en fonction de bornes supérieures sur le score.

Des domaines de motifs différents ont aussi été étudiés dans ce cadre du graphe dynamique tels que [CNB09, HC$^+$09, BBBG09, AK15]. [CNB09] ont par exemple reformulé le problème de fouille de graphes en problème de fouille d'un tenseur booléen (encore appelé fouille d'une relation n-aire). Le tenseur en question est de dimension trois et correspond à l'évolution de la matrice d'adjacence du graphe au cours du temps. Le graphe n'est ni étiqueté ni attribué. Les motifs recherchés sont des *3-sets* (T, V_1, V_2) correspondant à des sous-cubes des données vérifiant des contraintes (anti-)monotones (où T est un sous-ensemble de temps, et V_1, V_2 sont des sous-ensembles de noeuds du graphe). Les contraintes considérées sont la connexité, la contiguïté et la fermeture. Au final, les motifs recherchés correspondent à des noeuds formant des cliques maximales sur des temps consécutifs. Pour extraire ces motifs, les auteurs utilisent l'algorithme en profondeur *Data-Peeler* développé dans le cadre plus général de relations n-aires [CBRB08]. [HC$^+$09, BBBG09] étudient un problème relativement similaire à celui étudié dans [IW08], à savoir extraire les transformations intéressantes au sein d'un même graphe dynamique étiqueté. [HC$^+$09] proposent un nouveau domaine de motifs appelé règles de réécriture du graphe (*graph rewriting rules*) et se focalisent sur les motifs "compressant" au mieux les transformations d'après le principe de description de longueur minimale (*Minimum Description Length*). L'approche proposée est relativement naïve : extraction des sous-graphes communs, calcul des transformations, et compression de celles-ci de manière itérative (par agrégation de sous-graphes). [BBBG09] in-

troduisent un domaine de motifs différent directement dérivé des règles d'association (appelés *evolution rules*) et adaptent l'algorithme *gSpan* pour les extraire. [AK15] étudient encore un domaine de motifs différent. Ils recherchent des motifs co-évoluants (appelés *Coevolving Relational Motifs* ou CRM) dans un graphe dynamique étiqueté. Un CRM $(N, < M_1, ..., M_m >)$ correspond à l'ensemble des sous-graphes fréquents $< M_1, ..., M_m >$ associé à l'ensemble de noeuds N. Les sous-graphes M_i doivent avoir leurs noeuds inclus dans N et partager au moins une arête. De plus, ils doivent être composés d'un nombre minimum d'arêtes, de noeuds, et de sous-graphes fréquents. L'algorithme proposé pour extraire ces motifs effectue en parcours en profondeur de l'espace de recherche basé sur les algorithmes *prefixSpan* [PHMA+01] et *gSpan* [YH02].

Extraction de motifs dans des graphes dynamiques attribués La fouille d'un graphe dynamique attribué a encore peu été étudiée dans la littérature. Un grand nombre de données peuvent pourtant être modélisées sous cette forme (p.ex. les réseaux sociaux, les réseaux de transports ou plus généralement les données géographiques). Toutefois, cette problématique est complexe car elle combine à la fois l'explosion combinatoire de la fouille d'*itemsets*, de séquences et de graphes. La figure 2.6 présente un exemple de graphe dynamique attribué. Elle illustre l'évolution d'un graphe (noeuds, arêtes et attributs) sur trois temps (t_1, t_2, et t_3). Contrairement à l'exemple précédent, chaque noeud est associé à un ensemble d'informations. Dans le reste de ce manuscrit, nous noterons pour simplifier *abcd* l'ensemble de valeurs $\{a, b, c, d\}$.

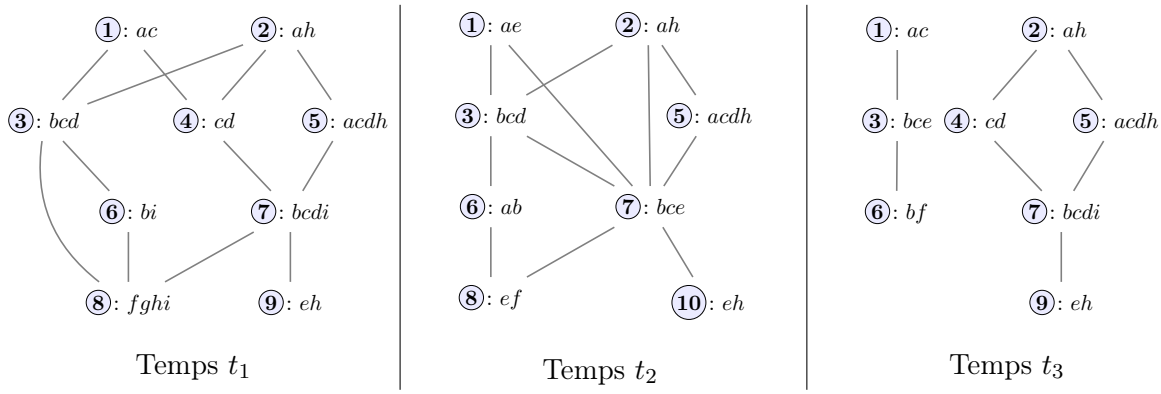


FIGURE 2.6 – Exemple de graphe dynamique attribué

Face à ces données complexes, [DPRB12] se concentrent sur l'extraction de co-evolution cohésives, et font l'hypothèse que les noeuds et les arêtes ne changent pas (seules les valeurs des attributs changent). Ces motifs représentent l'évolution d'un ensemble de noeuds voisins partageant les mêmes valeurs pour certains attributs pendant un ensemble de temps. Plus formellement, ces motifs corespondent à des triplets (N, T, P), où N est un sous-ensemble des noeuds du graphe, T est un sous-ensemble de temps (non nécessairement consécutifs) et P est un sous-ensemble de valeurs d'attributs. Les auteurs considèrent trois contraintes supplémentaires. Les motifs doivent être maximaux, i.e. le motif n'est plus solution si des noeuds ou des attributs sont ajoutés. Les arêtes associées aux noeuds du motif doivent être relativement similaires au cours du temps (d'après la mesure cosinus et la distance de Jaccard). Le volume du motif ($|N| \times |T| \times |P|$) doit aussi être suffisamment important. Ces motifs se rapprochent donc de *clusters* de noeuds extraits en fonction de sous-espaces des attributs et du temps (*subspace clustering*). Pour trouver ces motifs, les auteurs découpent l'espace de recherche en partitions indépendantes (pouvant tenir en mémoire) et ils les explorent en utilisant l'algorithme *Data-Peeler* développé pour extraire des motifs fermés dans des tenseurs booléens [CBRB08]. Ce travail est étendu dans [DPRB13] afin

d'intégrer des contraintes sur la structure topologique des noeuds et ainsi extraire des motifs moins fragmentés. Le diamètre des sous-graphes associés à un motif doit notamment être inférieur à un seuil défini par l'utilisateur. Les noeuds du motif doivent être suffisamment différents des autres noeuds (*vertex specificity*). La valeur d'un attribut du motif doit suivre une tendance suffisamment différente de la tendance générale (*trend relevancy*).

Dans le cadre de la fouille d'un graphe dynamique attribué, [KPPR15] définissent la notion de motif déclencheur (*triggering pattern*). Un motif déclencheur $< L, R >$ est une séquence de valeurs (séquence d'*itemsets*) L prises par des noeuds du graphe, avant une évolution topologique R. Par exemple, $< a^+b^+, a^-, deg^+ >$ représente une augmentation simultanée des attributs a et b suivi d'une diminution de l'attribut a, suivi d'une augmentation du degré des noeuds. N représente toujours une unique information topologique. Un grand nombre de contraintes sont utilisées pour filtrer les motifs intéressants. On retrouve la fréquence minimale du motif, son taux d'accroissement minimum par rapport à R (capacité de L à discriminer R) et sa maximalité (fermeture), mais également des contraintes sur la topologie des noeuds dans le graphe tels que leur couverture, leur degré, leur centralité, leur diamètre ou leur synchronicité. Pour extraire ces motifs, les auteurs transforment le graphe en collection de séquences. Chaque séquence correspond alors à l'évolution d'un noeud (valeurs d'attributs et topologie). Puis, ils utilisent l'algorithme *CloSpan* [YHA03] pour extraire des motifs séquentiels fermés fréquents. Afin d'améliorer l'efficacité des tests d'inclusion, ils intègrent aussi la table de hachage proposée dans [ZH02] pour la fouille d'*itemsets* fréquents.

Les travaux existants se concentrent donc sur des domaines de motifs très spécifiques et ne permettent pas d'extraire des motifs locaux plus généraux (p.ex. des évolutions de sous-graphes intéressants) comme cela a été fait pour la fouille de graphes "statiques".

Le tableau 2.4 résume les principaux travaux décrits précédemment.

TABLE 2.4 – Extraction de motifs dans des graphes : synthèse

Article	**Motifs extraits**	**Contraintes et Mesures**	**Méthode d'extraction**	**Applications**
Collection de graphes étiquetés				
[IWM00] (*AGM*)	graphes étiquetés induits	fréquence min.	stratégie d'*Apriori* [AS+94]	analyse de substances chimiques cancérigènes
[YH02] (*gSpan*)	graphes étiquetés induits et inclus	fréquence min.	stratégie de $PrefixSpan$ [PHMA+01] *canonical DFS code*	données synthétiques, données benchmark
[NK04] (*Gaston*)	chemins, arbres libres, graphes étiquetés	fréquence min.	parcours en profondeur basés sur les chemins puis les arbres fréquents	données benchmark
[WDW+08] (*DAGMA*)	DAG étiquetés induits	fréquence min.	stratégie d'*Apriori* [AS+94] et forme canonique	données synthétiques, optimisation de code logiciel
[FAL+13]	DAG étiquetés et pondéré	fréquence pondérée min.	stratégie de *DAGMA* [WDW+08]	données synthétiques
[DLLS15]	graphes étiquetés induits	fréquence min.	propagation de contraintes, stratégie de type *Apriori* ou *pattern growth*	analyse de composants chimiques et de protéines
Unique graphe étiqueté				
[CH94] (*SUBDUE*)	graphes étiquetés	description de longueur minimale pondérée, compression, connectivité, couverture	algorithme glouton basé sur une recherche en faisceau	analyse de composants chimiques , analyse d'un circuit électronique et analyse d'images

Suite sur la page suivante

TABLE 2.4 – *Suite de la page précédente*

Article	Motifs extraits	Contraintes et Mesures	Méthodes d'extraction	Applications
[KK05]	graphes étiquetés inclus	fréquence min., arêtes disjointes	parcours en largeur et parcours en profondeur, code séquentiel	données benchmark
[GSV06]	graphes étiquetés induits	fréquence min., connexité	stratégie d'*Apriori* [AS+94], extensions basée sur des chemins disjoints	données synthétiques, données benchmark
[LZY10] (*DESSIN*)	graphes étiquetés induits	fréquence min. et max. des noeuds	parcours en profondeur et structure d'indexation *GADDI* [ZLY09]	données synthétiques, données benchmark
Représentations condensées de graphes étiquetés fréquents				
[YH03] (*CloseGraph*)	graphes étiquetés induits	fréquence min., fermeture	stratégie de *gSpan* [YH02], *equivalent occurence*, et *early termination*	données synthétiques et analyse de composants chimiques
[TTN+07]	DAG inclus (étiquettes distinctes)	fréquence min., fermeture	algorithme *LCM* [UAUA04] et combinaisons d'arêtes, groupements ancêtres-descendants (*tiles*)	analyse de réseaux de gènes
[BHPG11]	graphe étiquetés	fréquence pondérée min., fermeture delta tolérante, fenêtre glissante	algorithme incrémental, représentation compressée, traitement par lot	données benchmark
Collection de graphes attribués				
[BB02] (*MoFa*)	graphes attribués inclus	fréquence min., constraste, connexité, taille min.	algorithme en profondeur de type *Eclat* [ZPOL97], génération de fragments en parallèle	analyse de molécules
[BMB04]	graphes attribués inclus	fréquence min., fermeture	stratégie d'*Eclat* [ZPOL97], *equivalent sibling pruning*, *perfect extension*	analyse de molécules
[CKK04]	DAG pyramidaux induits	fréquence min.,	stratégie de type *Apriori* [AS+94]	données synthétiques
[MOO09]	DAG avec *itemsets* quantitatifs	fréquence (nombre d'occurrences avec arêtes disjointes) min., densément connectés	algorithme *gSpan* [YH02] et *QFIMiner* [WMM05]	données benchmark
[MCRE09]	graphes attribués induits	seuil de cohésion de sous espace, densité min., connectivité, clique maximale	parcours par niveau, *expand-by-one*, fusion de candidats, table de hachage	analyse de réseaux de publications, analyse de gènes, données synthétiques
[FSKS10]	graphes attribués induits	*itemsets* partagés, maximaux, taille min.	parcours en profondeur, arbre préfixe, table des *itemsets* visités	réseaux biologiques, réseau de publications
[PFSSF17]	DAG attribués enracinés	fréquence min., connexité	algorithme en profondeur extraction d'arbres [PSFSF16], forme canonique	réseaux sociaux, usages site Web
Graphe dynamique étiqueté				
[LBW08]	graphes (sans étiquette)	fréquence min., périodicité min./max., fermeture	parcours en largeur, table de hachage, arbre de motifs	mails dans une entreprise, réseaux sociaux , communications mobiles
[BBBG09]	graphes temporels inclus et règles d'évolution	fréquence min., connexité, confiance min.	adaptation de *gSpan* [YH02]	données benchmark
[CNB09]	3-sets	connexité, contiguïté, fermeture	algorithme *DataPeeler* [CBRB08]	réseau de transports

Suite sur la page suivante

TABLE 2.4 – *Suite de la page précédente*

Article	**Motifs extraits**	**Contraintes et Mesures**	**Méthodes d'extraction**	**Applications**
[HC+09]	règles de réécriture du graphe	compression	approche naïve, description de longueur minimale	réseaux biologiques
[OO09]	séquences de graphes induits	fréquence min., taille min. corrélation min. [TKS02]	algorithme par niveau, arbres préfixe et postfixe	mails dans une entreprise, communications mobiles
[BMS11]	graphes (noeuds et arêtes fixes)	poids des arêtes max., connexité	heuristique, groupes de sous-intervalles de temps, borne sup.	réseau de transport, mails dans une entreprise réseau sociaux
[IW10]	séquences de graphes induits	fréquence min., connexité	algorithme *AGM* [IWM00] suivi de *PrefixSpan* [PHMA+01]	données synthétiques, données benchmark
[AK15]	motifs co-évoluants en relation	fréquence min., arête commune, nombre d'arêtes et de noeuds min.	adaptation de *gSpan* [YH02] et *PrefixSpan* [PHMA+01]	réseau de publications, réseaux biologique réseau de magasins
Graphe dynamique attribué				
[DPRB12]	co-évolution cohésives (noeuds et arêtes fixes)	*itemsets* identiques, maximaux, arêtes similaires, volume min.	algorithme *DataPeeler* [CBRB08], partitions indépendantes	données synthétiques, glissements de terrains réseau de publications
[DPRB13]	co-évolution cohésives (noeuds et arêtes fixes)	id. [DPRB13], topologie (diamètre spécificité des noeuds), tendance	algorithme *DataPeeler* [CBRB08]	réseau aérien glissements de terrains, réseau de publications
[KPPR15]	séquences déclencheuses	fréquence min., taux d'accroissement min., couverture min., degré min. centralité, diamètre min., synchronicité min., maximaux	parcours en profondeur basé sur *CloSpan* [YHA03], table de hachage [ZH02]	réseaux sociaux, réseau de publications, réseau aérien

2.2 L'extraction de motifs dans des données spatio-temporelles

Les sections suivantes vont introduire des contributions spécifiques aux données spatio-temporelles, avec un focus plus particulier sur les données "évènements" et "rasters". Une grande partie d'entre elles s'appuie directement sur les travaux effectués en fouille de séquences et de graphes, et présentés précédemment.

2.2.1 Suivi d'évènements ou d'objets spatiaux

L'extraction de motifs dans des bases de données spatiales d'événements a été au coeur de beaucoup de travaux [SEKM11]. Ils diffèrent principalement dans leur façon de définir les motifs et dans les contraintes spatio-temporelles utilisées.

Les co-localisations (*colocations*) sont des motifs particulièrement étudiés dans le cadre des données spatiales [HSX04, YS06, VAN07], et qui ont été étendus aux données spatio-temporelles par la suite. Ces motifs spatiaux sont des ensembles de types d'événements (ou d'objets) fréquemment associés à des objets spatiaux voisins (i.e. leurs instances forment des cliques). Le motif $\{mine, erosion, vegetation\_faible\}$ pourrait être un exemple de co-localisation associée à des zones où l'érosion serait étudiée. Il représenterait le fait que les objets *mine*,

erosion, et $vegetation\_faible$ sont souvent à proximité. Autrement dit, il y aurait une corrélation spatiale entre ces éléments. Les co-localisations spatio-temporelles ont été définies dans [CSRS06, CSRS08] comme une extension des co-localisations classiques afin de représenter des ensembles d'objets/événements voisins dans l'espace et dans le temps. Les instances de ces motifs sont spatialement proches pendant une fraction significative de temps. Ces contraintes ont été intégrées par l'intermédiaire d'une mesure d'intérêt monotone combinant prévalence spatiale et prévalence temporelle. Pour simplifier, seuls les motifs apparaissant un grand nombre de fois dans un grand nombre de temps sont conservés. Par exemple, dans la figure 2.7, les motifs $\{erosion, piste\}$ et $\{feux, vent, aireRepos\}$ sont des co-localisations spatio-temporelles (les traits en pointillés représentent la relation de voisinage). Toutefois, $\{feux, vent, aireRepos\}$ aura une mesure d'intérêt plus forte que $\{erosion, piste\}$ car il apparaît plus fréquemment dans un nombre de temps moins important ($\{erosion, piste\}$ n'apparaît qu'une seule fois à chaque temps). Pour extraire ces motifs, les auteurs ont utilisé une stratégie "générer-tester" et un parcours par niveau de l'espace de recherche de type *Apriori* [AS+94]. Les auteurs ont étendu leurs travaux dans [Cel15] afin d'intégrer une nouvelle contrainte temporelle sur la "durée de vie" de l'événement.

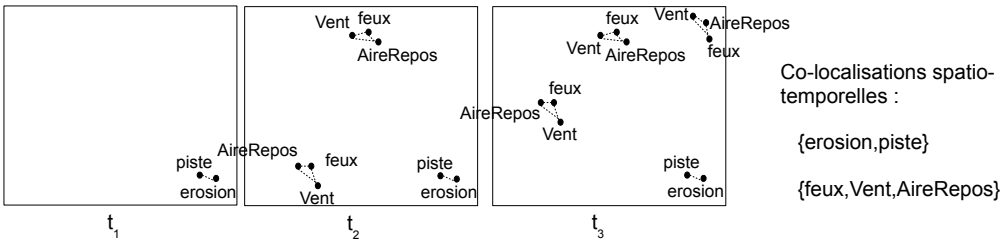


FIGURE 2.7 – Exemple de co-localisations spatio-temporelles

Le concept de co-localisations a aussi été étudié dans d'autres travaux tels que [YPM05, QHH09, PAA13]. Par exemple, [YPM05] proposent un cadre théorique pour l'extraction de motifs spatiaux apparaissant fréquemment à différents temps, et une extension permettant de visualiser certaines évolutions dans le temps. Ces motifs, appelés SOAP (*Spatial Object Association Pattern*), sont des ensembles fréquents de types d'objets/événements vérifiant des contraintes spatiales. Ce domaine de motifs est donc relativement similaire aux co-localisations. Il y a toutefois deux principales différences. Les calculs de distances et de voisinages sont basés sur les objets spatiaux et non des points. De plus, ces motifs peuvent représenter quatre types de configurations spatiales (clique, séquence, étoile, *minLink*) et non pas une seule. La figure 2.8 illustre les trois premières configurations. La dernière configuration permet de définir des SOAP plus généraux où seul le nombre minimum de voisins (*minLink*) associés à chaque objet est fixé. A noter que pour les séquences, les arêtes représentent une relation de voisinage et de direction. Par exemple, une arête (x, y) représente "x est voisin et au dessus de y". Une fois les SOAP fréquents extraits par un algorithme similaire à *Eclat* [ZPOL97], ils sont utilisés pour mettre en avant les évolutions de configurations d'un ensemble de types d'objets donné. Comme le montre la figure 2.8, l'évolution de l'ensemble $\{erosion, riviere, piste, foret\}$ est représentée par la séquence de SOAP fréquents associés à cet ensemble et apparaissant aux différents temps. Toutes les occurrences de SOAP fréquents ne sont pas considérées. Seules sont conservées les occurrences associées à un intervalle de temps où le motif est apparu puis a disparu.

Les travaux de Qian et al. dans [QHH09] se sont intéressés à l'extraction des SPCOZ (*Spread Patterns of spatio-temporal Co-occurrences Over Zones*). Ces motifs représentent la propaga-

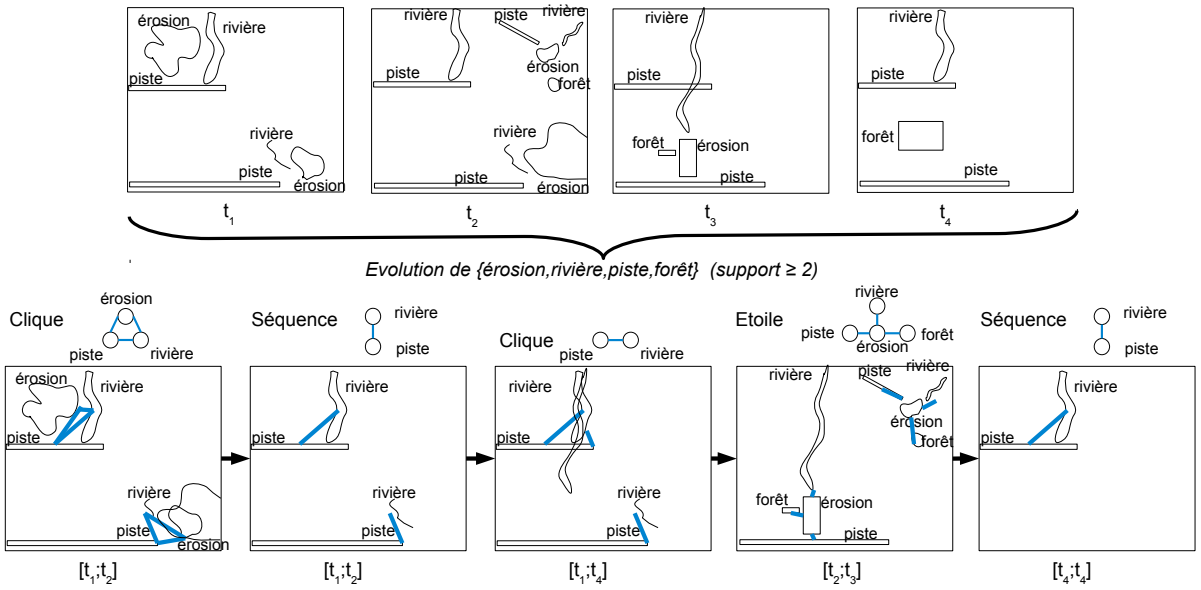


FIGURE 2.8 – Exemple de SOAP fréquents et visualisation de l'évolution d'un ensemble d'objets grâce à ces SOAP

tion (la "trajectoire") de co-localisations spatiales. Qian et al. ont ainsi étendu les travaux sur les co-localisations pour suivre le déplacement d'éléments de propagation (*spread element*). Un élément de propagation est une co-localisation "fréquente" localisée et associée à une fenêtre temporelle. Dans la figure 2.9, la co-localisation fréquente $\{erosion, piste\}$ associée à l'intervalle $[t_1, t_2]$ est un élément de propagation. Les éléments de propagation combinés deux à deux constituent des arbres représentant la propagation d'un motif (SP-Tree ou *Spread Pattern Tree*). La figure 2.9 illustre deux exemples de motifs SPCOZ : le SP-tree de $\{erosion, piste\}$ et celui de $\{feux, vent, aireRepos\}$. L'algorithme d'extraction développé commence par rechercher toutes les co-localisations fréquentes de taille 2. Puis, il les utilise pour générer les éléments de propagation et les SP-Tree correspondants. A cette étape, les motifs construits sont donc des arbres de propagation composés de co-localisations de taille 2. Les autres arbres de propagation (ceux ayant des co-localisations de taille supérieure à 2) sont obtenus par une approche par niveau de type *Apriori*, i.e. les arbres ayant des co-localisations de taille k sont obtenus à partir des sous-arbres composés des co-localisations de taille $k-1$.

Dans [PAA13], les auteurs proposent une nouvelle mesure basée sur l'air des objets étudiés pour filtrer les co-localisations intéressantes. Cette mesure peut être vue comme une adaptation de la distance de Jaccard à des objets spatiaux. Elle représente la surface relative partagée par toutes les instances d'une co-localisation. La mesure proposée a l'avantage d'être peu coûteuse, tout en garantissant certaines propriétés permettant d'élaguer l'espace de recherche pendant l'extraction. La contrainte associée (surface relative minimale) est notamment utilisée pour améliorer l'efficacité d'un algorithme de type *Apriori* [AS+94].

Les travaux précédents sont fortement liés aux travaux sur l'extraction d'*itemsets* fréquents. En effet, tous ces domaines de motifs sont des *itemsets* vérifiant des contraintes (anti-)monotones. Les algorithmes d'extraction sont donc directement dérivés de ceux définis dans ce contexte. De la même manière, certains travaux ont étendu les approches développées pour l'extraction de motifs séquentiels dans le contexte des données spatio-temporelles. Par exemple, Tsoukatos et al. dans [TG01] ont étendu les travaux sur les séquences d'*itemsets* afin d'extraire des séquences

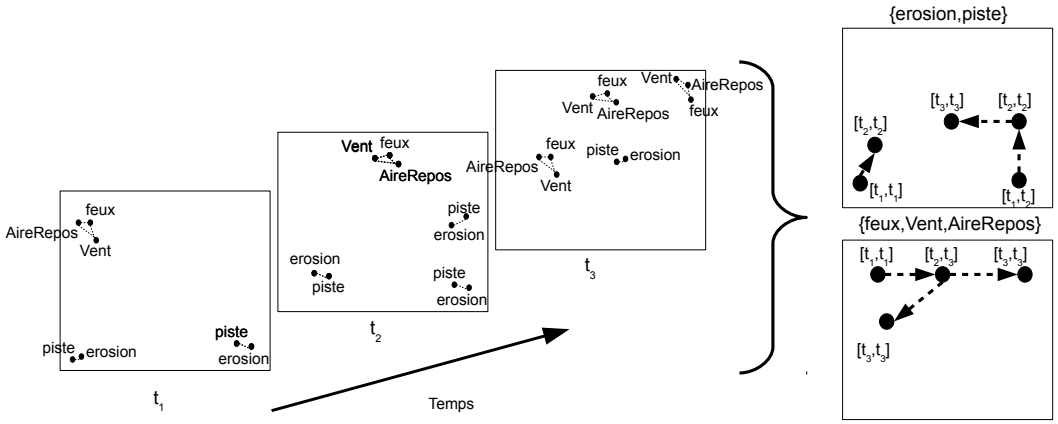


FIGURE 2.9 – Exemple de SPCOZ

représentant l'évolution dans le temps de zones d'études (p.ex. des quartiers). La base de données considérée est constituée de séquences d'*itemsets* représentant l'évolution temporelle des caractéristiques environnementales de différentes zones. Un algorithme effectuant un parcours en profondeur de l'espace de recherche est utilisé pour extraire les séquences les plus fréquentes (i.e. celles apparaissant dans le plus de zones). La figure 2.10 illustre un exemple de séquences pouvant être extraites. Les auteurs ont également proposé une approche pour extraire les séquences fréquentes à une granularité spatiale plus élevée (p.ex. région) en exploitant les séquences fréquentes trouvées à une granularité plus faible (p.ex. ville). Ils exploitent pour cela le fait que les séquences extraites à un niveau plus faible resteront fréquentes à un niveau de granularité plus élevé. La méthode recherche alors uniquement de nouvelles séquences fréquentes issues de l'agrégation spatiale.

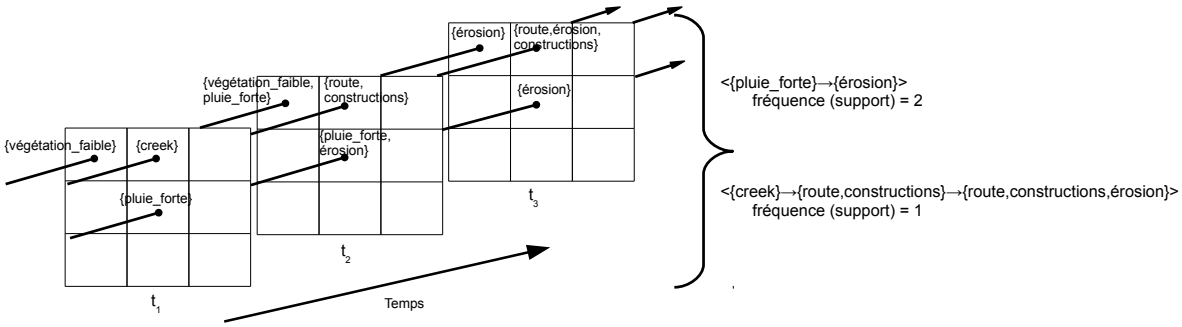


FIGURE 2.10 – Exemple de séquences représentant l'évolution de zones

Dans [WHLW04], les auteurs se focalisent sur l'extraction de séquences représentant la propagation spatio-temporelle d'événements dans des fenêtres temporelles prédéfinies. Dans cet objectif, ils découpent la dimension temporelle en fenêtres d'une taille donnée (p.ex. 4 jours), divisent l'espace sous la forme d'une grille, et introduisent le concept de *flow pattern*. Un *flow pattern* est une séquence d'ensembles d'événements de la forme $< E_1 \rightarrow E_2 \rightarrow ... \rightarrow E_k >$ où E_i est un ensemble d'événements de la forme $e(localisation)$, avec e un type d'événements (p.ex. pluie, vent). Chaque ensemble d'événements est composé d'événements spatialement voisins apparaissant au même temps. Deux ensembles d'événements E_p et E_q sont

consécutifs dans la séquence, si leurs événements appartiennent à la même fenêtre temporelle, s'ils sont tous voisins et qu'ils apparaissent à deux temps consécutifs. L'objectif de ce travail est de trouver les séquences d'événements apparaissant fréquemment. La figure 2.11 présente quelques *flow patterns* avec leur fréquence (dans cet exemple, deux événements sont voisins si leur distance euclidienne est inférieure ou égale à 1). Les événements sont localisés par des coordonnées (X,Y) tel que l'événement pluie(0,1) au temps t_1. A titre d'exemple, le flow pattern $< \{piste(0,0)\} \rightarrow \{erosion(1,0)\} >$ apparaît trois fois. A l'opposé, le motif $< \{piste(0,0), pluie(0,1)\} \rightarrow \{AirRepos(1,2), feux(1,2)\} >$ a une fréquence de zéro car les événements $\{AirRepos(1,2), feux(1,2)\}$ ne sont pas voisins de tous les événements de $\{piste(0,0), pluie(0,1)\}$ ($AirRepos$ et $feux$ du temps t_2 ne sont pas voisins de $piste$ du temps t_1). Il est aussi intéressant de noter que les motifs $< \{piste(0,0), pluie(0,1)\} \rightarrow \{erosion(1,0)\} >$ et $< \{piste(0,0), pluie(1,1)\} \rightarrow \{erosion(1,0)\} >$ sont considérés comme deux motifs différents bien que représentant des phénomènes similaires (seule la localisation de l'événement pluie est légèrement différente). Autre point important, la fréquence du motif $< \{AirRepos(1,2)\} \rightarrow \{AirRepos(1,2), feux(1,2)\} \rightarrow \{feux(2,2)\} \rightarrow \{vegetation_faible(1,2)\} >$ est égale à 0 car il n'est pas inclus dans une unique fenêtre temporelle. Pour extraire ces motifs, l'algorithme proposé suit une stratégie par niveau pour trouver les séquences de tailles un et deux, puis utilisent les motifs fréquents trouvés comme point de départ à un parcours en profondeur de l'espace de recherche. Dans un deuxième temps, Wang et al. [WHL05], étendent cette notion et définissent les motifs spatio-temporels généralisés (*generalized spatio-temporal pattern*) comme des séquences de *relative eventsets*. Un *relative eventset* est un ensemble d'événements dont la localisation est remplacée par un positionnement relatif à une localisation de référence. Pour extraire ces motifs, ils utilisent une approche dérivée de l'algorithme $PrefixSpan$ [PHMA+01].

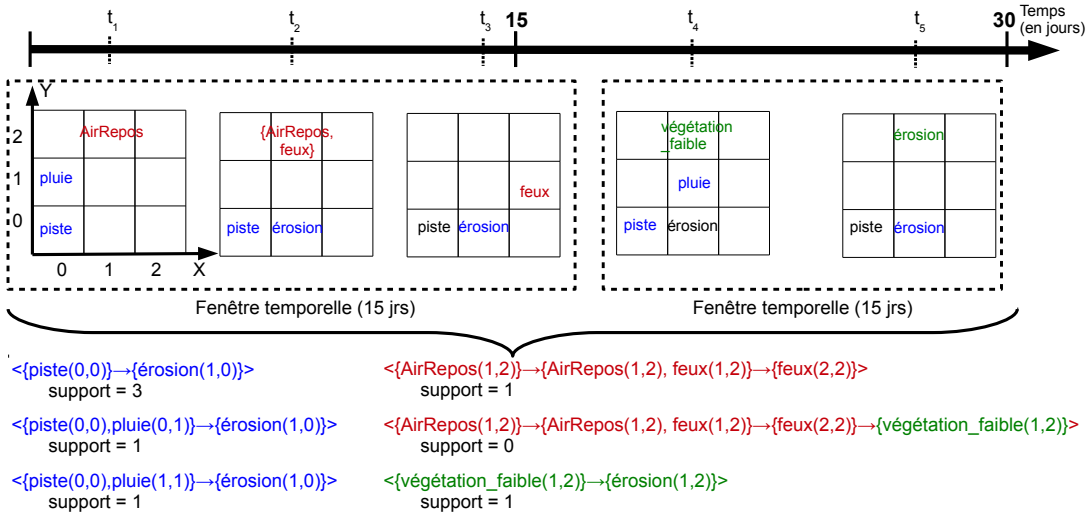


FIGURE 2.11 – Exemple de *flow patterns*

Huang et al., dans [HZZ08], se sont concentrés sur le problème d'extraction de séquences de propriétés représentant la propagation de certains types d'événements. Ces séquences sont de la forme $< f_1 \rightarrow f_2 \rightarrow ... \rightarrow f_k >$, où f_i est un type d'événements (et non un ensemble). Cette approche permet donc d'étudier la propagation des événements pris individuellement (sans prendre en compte leur environnement). Ce modèle considère deux événements consécutifs s'ils sont spatialement proches (distance euclidienne inférieure à un seuil donné) et apparaissent dans la même fenêtre temporelle. Les auteurs ont également étudié d'autres relations de voisinage

dépendant du temps. Ces relations permettent de représenter un rétrécissement de la zone d'influence d'un événement (son voisinage) au fur et à mesure que le temps passe (c'est ce qui se passe lors de la propagation d'une maladie infectieuse). Les auteurs proposent aussi une nouvelle mesure d'intérêt pour ces séquences car pour eux, les mesures basées sur la fréquence ne reflètent pas nécessairement un lien de cause à effet entre les événements. Leur mesure, appelée *sequence index*, s'appuie sur des travaux antérieurs [Cre93] exploitant des statistiques spatiales pour étudier l'indépendance de phénomènes. Ainsi, une séquence est intéressante si le ratio de densité spatiale entre deux éléments consécutifs de cette séquence est supérieur à un seuil. Ce ratio de densité est proche de 1 si les deux types d'événements sont distribués de manière indépendante. Plus la mesure est inférieure à 1, plus les types d'événements se "repoussent". Au contraire, plus ce ratio est au dessus de un, plus les chances d'avoir un lien de cause à effet entre ceux-ci sont élevées. La principale limite de cette contrainte est de ne pas être anti-monotone. Elle ne peut donc pas être utilisée directement pour élaguer l'espace de recherche. Pour extraire ces séquences, les auteurs proposent donc un nouvel algorithme *Slicing-STS-Miner* basé sur un traitement incrémental des différentes fenêtres temporelles et une extension des séquences à chaque étape. Au lieu de s'appuyer sur l'anti-monotonie, ils exploitent une autre propriété du *sequence index* : si une séquence est intéressante, toutes les sous-séquences ayant le même préfixe sont intéressantes.

[Pat10] utilise la durée des événements pour modéliser l'influence temporelle d'un événement sur un autre. Il propose un nouveau domaine de motifs séquentiels, appelés *temporal-spatial feature interaction patterns*, représentant des séquences de types d'événements. En plus des types d'événements, ces motifs contiennent des informations sur leur direction (nord, nord-est, est, etc) ainsi que des informations sur les relations temporelles entre les événements (p.ex. rencontre, recouvre, contient, avant, etc). L'approche proposée prend en compte huit types de relations temporelles. L'algorithme développé permet d'extraire incrémentalement l'ensemble des séquences fréquentes à partir des motifs fréquents de taille deux. Pour améliorer l'efficacité de leur extraction, l'auteur s'appuie sur un découpage de l'espace et du temps, ainsi qu'une structure de données arborescente pour stocker les motifs.

Mohan et al. dans [MSSR12] étendent aussi les travaux de [HZZ08] en étudiant des graphes orientés acycliques de types d'événements (un DAG étiqueté). Ces graphes représentent les événements voisins dans l'espace et le temps. Plus précisément, deux types d'événements f_1 et f_2 sont liés par une arête orientée de f_1 vers f_2, si f_2 apparaît peu de temps après f_1 à une localisation proche. Leur objectif est d'extraire des sous-DAG fréquents, appelés motifs spatio-temporels en cascade. Tout comme précédemment, ces motifs ne considèrent pas l'environnement proche d'un événement à un temps donné. Les auteurs introduisent une nouvelle mesure d'intérêt appelée *cascade participation index* construite à partir de la mesure proposée dans [HSX04]. Grâce à la monotonie de cette mesure, un algorithme par niveau de type *Apriori* [AS$^+$94] a été développé pour extraire les motifs les plus intéressants.

Dernièrement, [BTD17] proposent un algorithme "non-paramétrique" pour extraire des séquences de taille deux de types d'événements (p.ex. $f_1 \rightarrow f_2$). Aucun seuil ni relation de voisinage n'est passé en paramètre. Ce travail s'appuie sur une modélisation des relations deux à deux entre événements suivant un modèle de Hawkes spatio-temporel multivarié (une classe de processus stochastiques ponctuels mutuellement excités). Ainsi, la probabilité qu'un type d'événements en précède un autre est utilisée comme mesure d'intérêt. Elle dépend à la fois des probabilités de cause à effet entre pairs d'événements et des probabilités que les événements soient aléatoires. Un classement des motifs est effectué pour faire ressortir les séquences intéressantes. L'extraction des motifs est faite selon une approche, appelée *stochastic declustering*, développée au départ pour analyser des séismes. Son objectif initial est de séparer les séismes principaux des répliques en estimant la probabilité qu'un séisme soit lié à un précédent séisme.

Le tableau 2.5 résume les travaux décrits précédemment.

TABLE 2.5 – Extraction de motifs dans des données événements : synthèse

Article	**Motifs extraits**	**Contraintes et Mesures**	**Méthode d'extraction**	**Applications**
[TG01]	séquences d'ensembles de types d'événements (séquences d'*itemsets*)	fréquence min., zones fixes	parcours en profondeur, agrégations à différentes granularités spatiales	données synthétiques
[WHLW04]	*flow patterns* (séquences d'ensembles d'événements)	fréquence (temps) min., proximité spatiale, événements consécutifs, fenêtres temporelles	parcours en profondeur, *vertical bitmap*, arbre de hachage	données synthétiques, météorologie, feux de forêts
[YPM05]	évolutions d'objets spatiaux *SOAP* (clique/séquence/étoile /*minLink*)	fréquence min., proximité spatiale, nombre d'occurrences min. à la fin, topologie	stratégie d'*Eclat* [ZPOL97] (représentation verticale, classe d'équivalence), *minimum bounding box*	simulation de molécules, simulation de vortex, données benchmark
[CSRS08]	ensembles de types d'événements (co-localisation)	indice spatial de participation min., persistence temporelle min., clique spatiale	stratégie d'*Apriori* [AS$^+$94], table de persistence temporelle, élagage temporel	données synthétiques, données benchmark
[HZZ08]	séquences de types d'événements (séquences d'*items*)	ratio de densité min., proximité spatio-temporelle	parcours en profondeur, voisinage dynamique, projections temporelles	données synthétiques, climat
[QHH09]	ensembles de types d'événements (co-localisation) et arbres de propagation (SPCOZ)	indice spatial de participation min., temps continus max., clique spatiale,	parcours par niveau	données synthétiques, activité commerciale
[Pat10]	*temporal-spatial feature interaction patterns* (séquences d'*items* avec des directions)	indice spatio-temporel de participation min., proximité spatiale et temporelle	parcours par niveau incrémental, arbre de motifs, partitionnement	inondations, végétation, vidéos sur l'activité sportive
[MSSR12]	graphes orientés acycliques de types d'événements (DAG étiquetés)	indice de participation en cascade min., proximité spatiale et temporelle	stratégie d'*Apriori* [AS$^+$94], filtrages par borne sup. et multi-résolutions	crimes
[PAA13]	ensembles de types d'événements (co-localisation)	indice spatial de participation min., coefficient de co-occurrence min. (volume), clique spatiale	stratégie d'*Apriori* [AS$^+$94], analyse basée sur la mesure de Jaccard, *filter-and-refine*	activité solaire
[Cel15]	ensembles de types d'événements (co-localisation)	indice spatial de participation min., indice temporel de participation min.	stratégie de [CSRS08]	données synthétiques, données benchmark
[BTD17]	séquences de types d'événements (séquences d'*items*)	probabilité min. entre paires d'événements	*stochastic declustering* [ML08] suivant un modèle de Hawkes	données synthétiques, accidents

2.2.2 Analyse d'une série temporelle de rasters

On peut distinguer trois types de travaux visant à extraire des motifs dans des données rasters : les travaux centrés sur la dimension spatiale, ceux centrés sur la dimension temporelle, et ceux intégrant les deux. Tout comme précédemment, les approches développées s'appuient fortement sur les travaux réalisés en fouille d'*itemsets*, de séquences et de graphes.

Approches centrées sur la dimension spatiale [KHA98] discutent des premiers travaux visant à faire de la fouille de données dans des rasters issus de données géographiques. Ces travaux se focalisent sur la dimension spatiale et intègrent peu la dimension temporelle. Les données en entrée sont des collections d'images d'astronomie, et l'objectif est principalement d'extraire des descripteurs pour faire de la classification supervisée.

De manière générale, l'analyse spatiale des données rasters peut se situer à deux niveaux : au niveau pixels et au niveau objets. Initialement, ces deux approches sont associées au traitement d'images. Toutefois, elles peuvent aussi s'appliquer plus généralement à tout type de données rasters. Dans les approches orientées pixels, l'unité d'analyse est le pixel. Autrement dit, les méthodes d'analyse considèrent en entrée un ensemble de pixels, et étudient leurs évolutions en prenant en compte certaines fois d'autres pixels (p.ex. voisins). Dans les approches orientées objets, l'unité d'analyse est l'objet, i.e. un ensemble de pixels homogènes. Il est donc nécessaire de détecter ces objets et de les associer dans le temps, avant de commencer l'analyse. La détection est classiquement faite par des méthodes de *clustering* (ou de segmentation). Un intérêt des méthodes orientées objets est de limiter l'influence du bruit. En effet, les pixels proches (spatialement et en valeur) sont regroupés et leurs caractéristiques sont affectées à l'objet.

Dans la littérature, l'approche objets et l'extraction de motifs ont beaucoup été utilisées en analyse d'images, notamment pour caractériser des collections d'images [DPDR01, NTU$^+$07, YWY07, FFT12, VCJ14, RFDT15, LLSvdH15, PS15, GB17, LXS17]. Par exemple, [YWY07] utilisent des co-localisations fréquentes pour cela. Tout d'abord, ils détectent des objets représentatifs (des *clusters* de pixels) dans les images puis les associent à des primitives visuelles. Chaque image est ainsi caractérisée par un sac de mots visuels (*bag-of-visual-words*). Les co-localisations fréquentes sont extraites à partir de l'ensemble de ces sacs par l'algorithme *FP-growth* [GZ03]. [FFT12] utilisent une approche similaire pour faire de la classification d'images. Dans ce travail, les sacs de mots visuels sont toutefois associés à des histogrammes. Ainsi, les motifs recherchés sont des ensembles fréquents d'histogrammes locaux. Les auteurs utilisent l'algorithme en profondeur *LCM* [UAUA03] pour extraire ces motifs, puis ils les filtrent pour ne conserver que les plus discriminants et représentatifs. [RFDT15] cherchent à résumer et à faciliter l'exploration d'une collection d'images. Leur idée est d'utiliser les motifs fréquents pour caractériser les images, puis de proposer une navigation entre les images à partir de ces motifs. Pour cela, ils extraient tout d'abord dans chaque image un ensemble de groupes de pixels discriminants (*mid-level clusters*) [SGE12]. Puis, ils recherchent les ensembles de *clusters* fréquents dans la collection grâce à l'algorithme *LCM* [UAUA03]. Beaucoup d'autres travaux en analyse d'images utilisent les *itemsets* fréquents tels que [DPDR01, NTU$^+$07, VCJ14, LLSvdH15, PS15, GB17, LXS17]. Leur prise en compte de la dimension spatiale se limite donc généralement à la détection des objets et leur caractérisation par un ensemble de valeurs.

Plusieurs techniques ont été développées pour représenter l'image sous la forme d'un graphe. On retrouve des représentations hiérarchiques tels que les *quad-trees* [FB74] ou des représentations basées sur le voisinage tels que les graphes d'adjacence de régions (*regions adjacency graph*). Ces graphes représentent des objets détectés (p.ex. par segmentation) ou des sous-régions de l'image vérifiant certaines propriétés. Ils sont au centre d'un grand nombre de méthodes en analyse d'images [LG12]. Un certain nombre de travaux ont utilisé des techniques de fouille de graphes fréquents dans ce contexte [JC09, OA10, AMGAMP12, AMGACO$^+$16, DFJS16]. Par exemple, [JC09] étudient un problème de classification à partir d'une collection d'images. Chaque image est transformée en arbre représentant des sous-régions homogènes (*quad-tree*) qui sont étiquetées de manière supervisée (noeuds et arêtes). Les sous-graphes fréquents sont utilisés comme descripteurs des images pour la classification. Ils sont extraits par l'algorithme *gSpan* [YH02] adapté pour prendre en compte le poids des arêtes. Ces poids dépendent de la fréquence relative

de l'arête dans chaque graphe en entrée. [ECGFS10] comparent cette approche avec une approche basée sur une représentation moins fine de l'information spatiale (représentation séquentielle). Les résultats obtenus mettent en avant l'intérêt des sous-graphes pour décrire des images et les classer. Le travail de [OA10] se place dans le même contexte. Les auteurs transforment chaque image en graphe étiqueté de régions d'intérêt obtenues par la méthode MSER (*Maximally Stable Extremal Regions*) [MCUP04]. Les arêtes du graphe représentent la proximité spatiale des régions (dérivée du diagramme de Voronoï), et les étiquettes correspondent à des identifiants de *clusters* obtenus à partir de leurs caractéristiques (morphologie, granulométrie, valeurs, etc). Ces graphes sont ensuite parcourus à la recherche de sous-graphes fréquents. La mesure de fréquence utilisée dans ce travail est basée sur le nombre de plongements du motif (*embeddings*) dans les données comme défini dans [FB07]. Les motifs trouvés sont utilisés pour caractériser chaque image en entrée d'un classifieur SVM multi-classes [Vap63]. [AMGAMP12] utilisent des sous-graphes fréquents approximatifs comme descripteurs pour une classification d'images basée sur SVM [Vap63]. Chaque image est découpée récursivement en quadrants en fonction des valeurs des pixels, et représentée sous la forme d'un graphe étiqueté (p.ex. par un indice de couleur). La collection de graphes ainsi constituée est ensuite fouillée d'après une adaptation de l'algorithme APGM (APproximate Graph Mining) [JZH11]. Cet algorithme effectue un parcours en profondeur de l'espace de recherche, mais utilise un test approximatif d'isomorphisme (basé sur des probabilités de substitution) au lieu du test classique. Ce test a l'avantage d'identifier une sous-structure même si elle n'apparait pas sous une forme ou une orientation identique. Ils étendent leur travail dans [AMGACO$^+$16] pour extraire des sous-graphes approximatifs émergeants, et ainsi mieux classer les images.

Approches centrées sur la dimension temporelle Une série temporelle de rasters peut aussi être vue comme un ensemble de séquences ou de séries temporelles spatiales (*spatial time series*) décrivant l'évolution de valeurs qui sont souvent numériques. L'extraction de motifs dans des séries temporelles a été très étudiée dans la littérature [Mue14, TL17]. Les travaux s'articulent principalement autour de deux familles de problèmes : la recherche de motifs fréquents et la recherche de motifs similaires. La première famille recherche des sous-séquences apparaissant de manière répétée dans la série temporelle. Elle est généralement associée à la recherche de motifs (séquentiels) locaux et de règles d'association. La deuxième famille recherche des segments (ou fragments) ayant une faible distance, i.e. une forte similarité, entre eux. Elle est généralement associée à la recherche de motifs globaux et à la classification. Certains de ces travaux étudient des séries temporelles de rasters [PGMF10, PIG12, RdAC$^+$13, PIV$^+$15], sans toutefois prendre en compte la dimension spatiale (ou très partiellement).

Par exemple, le travail présenté dans [PGMF10] étudie la détection de changements dans une série temporelle d'images satellitaires. La détection de changements est habituellement effectuée en comparant les images deux à deux, au niveau pixel, et en analysant leurs différences [LMBM04]. Ces comparaisons ne considèrent donc pas les changements ayant lieu sur de longues périodes de temps ou cycliques. Face à cette limite, les auteurs proposent une approche différente. Ils extraient les séquences fréquentes (maximales) d'ensemble de valeurs de pixels. Ils prennent donc en compte l'ensemble du vecteur de valeurs associé au pixel. Deux autres contraintes sont ajoutées lors de l'extraction afin de cibler les motifs représentant des changements. Ils élaguent les motifs de taille 1 qui sont trop fréquents ainsi que les motifs ayant deux valeurs identiques successives pour une même bande. Pour l'extraction, les auteurs utilisent l'algorithme par niveau PSP [MCP98].

Par la suite, [PIG12] recherchent des motifs dans une série temporelle d'images satellitaires

associée à une même zone urbaine. L'objectif est de les utiliser pour effectuer une classification de la couverture terrestre. Ils se focalisent donc sur la définition d'une mesure de similarité permettant de regrouper les pixels en prenant en compte leur évolution dans le temps. Chaque pixel est associé à une série temporelle multidimensionnelle de valeurs radiométriques. En raison de la présence de nuages et d'événements avec des cycles variés, ces séries temporelles sont très irrégulières. Elles ont des longueurs et des échantillonnages dans le temps différents. Ils adaptent donc la mesure de distance *DTW* (*Dynamic Time Warping*) [SC78] pour prendre en compte ces irrégularités. *DTW* est une méthode classique pour calculer un appariement optimal entre deux séries temporelles sous certaines contraintes. Comme une grande partie des travaux étudiant ce type de motifs "similaires", ils utilisent des algorithmes de classification standards (*k-means* [Llo82] et *clustering* hiérarchique ascendant) et se focalisent sur la définition de la mesure de similarité permettant de comparer les séries.

[RdAC+13] étudient l'extraction de motifs dans des séries temporelles de données rasters liées à l'agriculture et au climat. Dans leur travail, chaque pixel est aussi associé à des séries temporelles de valeurs (une pour chaque bande/attribut du raster). Ils définissent trois types de segments en fonction de l'évolution des valeurs dans le temps et caractérisent les séries temporelles en fonction de ces segments. Par exemple, le segment V (*Valley*) représente une diminution suivie d'une augmentation des valeurs, et le segment M (*Mountain*) représente l'inverse. Ils transforment ainsi les séries temporelles de valeurs numériques en séries symboliques (i.e. des séquences d'*items*). Une fois les séries symboliques générées, des associations de tendances fréquentes sont recherchées entre les séries, et des règles sont générées (p.ex. $Pluie[M] \rightarrow Temperature[V]$). Ces associations se limitent ici à des ensembles de deux tendances et à des règles de type $A \rightarrow B$. Les mesures de fréquence et de confiance sont très similaires à celles classiquement utilisées. Elles considèrent seulement la longueur des séries à la place de leur nombre et peuvent intégrer une fenêtre temporelle (glissante) dans leur calcul.

L'objectif dans [PIV+15] est de prédire des types de productions agricoles dans une série temporelle de rasters. Les images satellitaires disponibles à chaque temps sont décomposées en objets (des parcelles agricoles) en fonction d'informations collectées sur le terrain. Chaque objet est caractérisé par différentes informations (attributs) issus des pixels de l'image (p.ex. indices de végétation et de texture) et de données collectées sur le terrain (p.ex. type de sol, quantité de pluie, ou nom du village). La série temporelle est ainsi transformée en une collection de séquences, représentant chacune l'évolution des attributs (discrétisés) d'une parcelle. Des motifs séquentiels multidimensionnels fréquents sont ensuite recherchés afin de caractériser les différents types de parcelles. Ces motifs sont des séquences de sous-ensembles de valeurs (les valeurs possibles étant limitées par les différents attributs considérés). L'extraction de ces motifs est effectuée par l'algorithme M^2SP proposé dans [PCL+05], et qui est une adaptation de l'algorithme par niveau PSP [MCP98]. Des règles temporelles sont ensuite construites à partir de ces motifs et utilisées pour faire de la prédiction.

Approches intégrant dimensions spatiales et temporelles Des travaux ont intégré la dimension spatiale dans leur analyse des séries temporelles, permettant ainsi de prendre en compte les auto-corrélations spatiales entre les séries. Certains de ces travaux s'appuient sur des approches existantes qu'ils adaptent en intégrant par exemple de nouvelles contraintes.

Par exemple, [HK01] proposent une approche orientée objets pour extraire des épisodes fréquents (encore appelés règles d'association temporelle dans l'article) dans une série d'images satellitaires. Les auteurs utilisent les cartes de Kohonen (*self-organizing map* ou SOM) pour identifier les objets dans chaque raster (i.e. les regroupements relativement homogènes de pixels).

Afin de pouvoir identifier un objet même si celui-ci se déplace, une méthode en deux étapes est mise en place. La première étape divise toutes les images selon une même grille et génère une SOM regroupant les cellules ayant des valeurs proches. Ces *clusters* représentent donc les différents types de cellules indépendamment de leur position et des images. Ils sont caractérisés par leur histogramme de fréquence et sont étiquetés sémantiquement par les experts (p.ex. typhon, anticyclone). La deuxième étape projette ces *clusters* dans les images d'origine, puis les regroupe en fonction de leurs histogrammes via une deuxième SOM. Une fois les différents objets détectés dans cette dernière SOM, la série temporelle d'images est transformée en une collection de séquences d'identifiants de *clusters* (i.e. d'objets). Les motifs sont générés à partir de cette collection. Ils correspondent à des épisodes fréquents en série (*serial episode*) tels que définis dans [MTV97a], avec une contrainte supplémentaire de cohésion entre événements. Un épisode est une séquence d'événements fréquents dans une fenêtre temporelle prédéfinie (les fenêtres sont glissantes). Dans ce travail, un événement est un tuple composé de l'identifiant d'un *cluster*, de sa durée de vie, de sa date de début et de sa date de fin. Les événements doivent avoir deux à deux une cohésion minimale. Cette mesure est dérivée d'une mesure utilisée en fouille de texte pour mesurer la co-occurence de deux termes. Leur extraction est effectuée en suivant une stratégie par niveau de type *Apriori* [AS+94]. La figure 2.12 illustre ce processus.

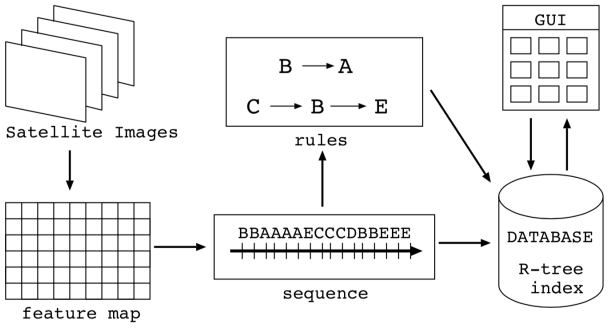


FIGURE 2.12 – Exemple d'extraction de règles temporelles dans une série d'images satellitaires [HK01]

D'autres travaux comme [JMB+11] ont aussi utilisé les motifs séquentiels pour rechercher des évolutions fréquentes de pixels dans une série d'images satellitaires. En plus d'être fréquentes, ces évolutions doivent avoir une surface et une connectivité minimales. Ils définissent pour cela un nouveau domaine de motifs, appelés *Grouped Frequent Sequential (GFS) pattern*. Ces motifs sont des sous-séquences de valeurs de pixels qui apparaissent dans suffisamment de pixels. Autrement dit, les motifs extraits doivent couvrir une surface minimum. La figure 2.13 présente un exemple de motif extrait dans une série de rasters. Comme le montre la figure, les motifs extraits sont assimilables à des séquences d'*items*. Dans leur travail, les auteurs ne considèrent donc pas tout le vecteur de valeurs associé à chaque pixel (i.e. toutes les bandes de l'image). Ils se focalisent uniquement sur l'indice de végétation (le NDVI, *Normalized Vegetation Index*) car ils étudient l'évolution de surfaces agricoles. De plus, en moyenne, chaque occurence du motif doit être voisine d'un nombre minimum d'autres occurrences (selon la méthode des huit plus proches voisins). Ces deux contraintes étant monotones, les auteurs adaptent l'algorithme $PrefixSpan$ [PHMA+01] afin de prendre en compte en plus la contrainte de connectivité minimum. Le nombre de motifs extraits peut être très important, ce qui pose notamment des problèmes d'interprétation. Face à ce problème, les auteurs étendent leur travail dans [MRP15] afin de mettre en avant les motifs les plus intéressants (sans faire d'hypothèse a priori). Pour cela, ils adaptent une méthode d'essai

randomisé (*randomization testing method*) pour vérifier si le motif aurait pu être observé dans un jeu de données aléatoire. Une série "aléatoire" de rasters est générée en échangeant aléatoirement la position des pixels et la "chronologie" des valeurs. La fréquence globale des valeurs de pixels ne change donc pas. L'information mutuelle normalisée (*normalized mutual information*) du motif dans les données initiales et dans les données aléatoires est calculée, puis elle est utilisée pour classer les motifs.

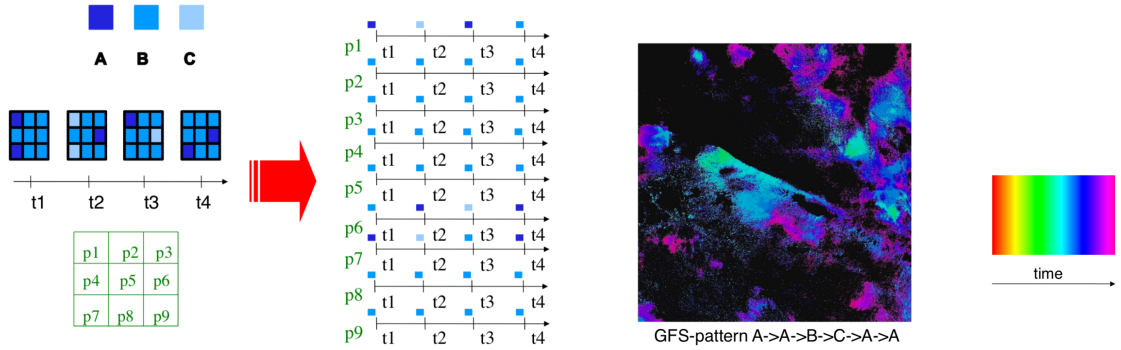


FIGURE 2.13 – Exemple de motif GFS extrait dans une série d'images satellitaires [MRP15]

Contrairement aux travaux précédents, [PJFD13] proposent un nouveau domaine de motifs de type graphe, et des algorithmes pour les extraire. Leur objectif est de suivre des objets fréquents dans une vidéo. Dans ce travail, la vidéo est vue comme une série temporelle d'images dans laquelle il faut extraire des motifs fréquents. Pour cela, les auteurs transforment chaque image en graphe planaire (*planar graph*) où les noeuds représentent les régions de l'image et les arêtes représentent les relations d'adjacence entre ces régions. Les régions de l'image sont détectées par segmentation en fonction de la valeur de leurs pixels. Les barycentres de ces régions deviennent des

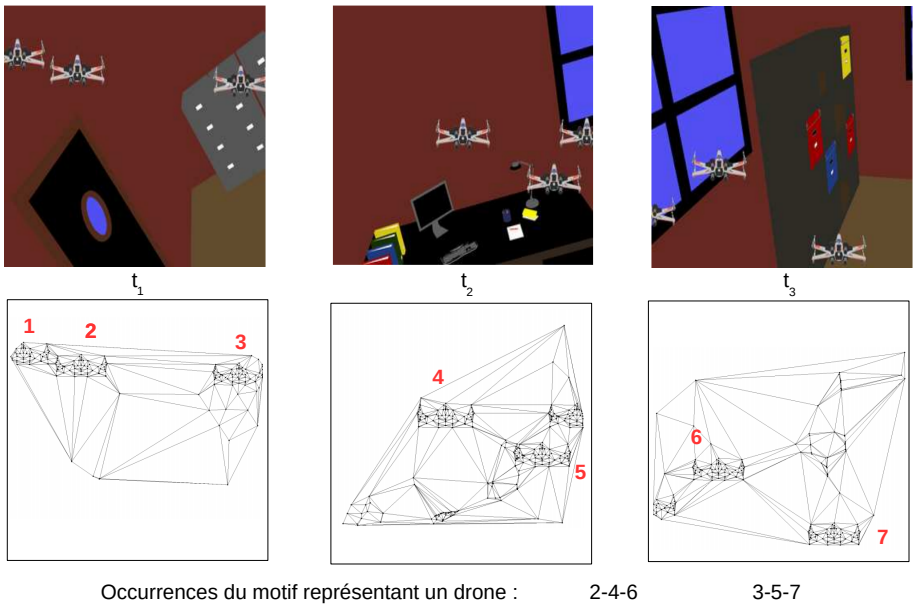


FIGURE 2.14 – Exemple de motifs de type graphe planaire extraits dans une vidéo [PJFD13]

noeuds, et les étiquettes des noeuds correspondent aux tailles des régions associées (discrétisées). La vidéo est ainsi transformée en série de graphes planaires étiquetés. La figure 2.14 illustre cette transformation pour une vidéo suivant des drones. Ensuite, l'objectif est d'extraire des sous-graphes (connexes) planaires fréquents apparaissant à proximité dans l'espace et le temps. Face à ce problème, les auteurs proposent un algorithme dérivé de *gSpan* [YH02] effectuant un parcours en profondeur de l'espace de recherche. Il s'appuie principalement sur certaines propriétés des extensions afin de limiter le nombre de motifs générés et sur une représentation des graphes sous forme de séquences d'arêtes (*canonical code*) afin limiter le coût du test d'isomorphisme.

Le tableau 2.6 résume les travaux décrits précédemment.

TABLE 2.6 – Extraction de motifs dans des rasters : synthèse

Article	**Motifs extraits**	**Contraintes et Mesures**	**Méthode d'extraction**	**Applications**
		SPATIAL		
[YWY07]	ensembles de types d'événements (co-localisations)	fréquence min., proximité spatiale	algorithme *FP-growth* [GZ03]	description de visages et de voitures
[FFT12]	ensemble d'histogrammes d'objets (*itemsets*)	fréquence min., proximité spatiale	algorithme *LCM* [UAUA03]	données benchmark en classification d'images
[RFDT15]	ensemble d'objets (*itemsets*)	fréquence min., proximité spatiale	algorithme *LCM* [UAUA03]	navigation dans une collection d'images
[JC09]	sous-graphes pondérés d'objets étiquetés	fréquence min., poids min., proximité spatiale	stratégie de *gSpan* [YH02]	données synthétiques
[OA10]	sous-graphes d'objets étiquetés	nombre d'occurrences (*embeddings*) min., proximité spatiale	algorithme *MoFa* [BB02]	détection d'habitats urbains
[AMGAMP12]	sous-graphes d'objets étiquetés	fréquence min. (similarité approximative), proximité spatiale	algorithme *APGM* [JZH11]	données synthétiques
		TEMPOREL		
[PGMF10]	séquences d'ensemble de valeurs de pixels (séquences d'*itemsets*)	fréquence min., maximaux	algorithme *PSP* [MCP98]	suivi environnemental
[PIG12]	segments de valeurs (*time series*)	similarité *DTW* multidimensionnelle irrégulière	*k-means* [Llo82] et *clustering* hiérarchique ascendant	suivi environnemental
[RdAC+13]	paires de tendances d'attributs	fréquence min., confiance min., fenêtre temporelle, changements	approche naïve (générer-tester toutes les combinaisons)	suivi de parcelles agricoles et suivi climatique
[PIV+15]	séquences multidimensionnelles de valeurs de pixels (séquences d'*itemsets*)	fréquence min.,	algorithme M^2SP [PCL+05] (adaptation de *PSP* [MCP98])	suivi de pratiques agricoles
		SPATIO-TEMPOREL		
[HK01]	séquences d'objets (séquences d'*items*)	fréquence min., fenêtre temporelle et cohésion min.	stratégie d'*Apriori* [AS+94]	suivi de typhons
[JMB+11]	séquences de valeurs de pixels (séquences d'*items*)	fréquence min. et connectivité	stratégie de *PrefixSpan* [PHMA+01]	suivi de pratiques agricoles
[MRP15]	séquences de valeurs de pixels (séquences d'*items*)	fréquence min., connectivité, information mutuelle normalisée	stratégie de *PrefixSpan* [PHMA+01] et méthode d'essai randomisé	suivi environnemental

Suite sur la page suivante

TABLE 2.6 – *Suite de la page précédente*

Article	Motifs extraits	Contraintes et Mesures	Méthodes d'extraction	Applications
[PJFD13])	sous-graphes planaires d'objets	fréquence min. et connectivité/adjacence	stratégie de *gSpan* [YH02]	suivi d'objets dans des vidéos

2.3 Positionnement des contributions

Les sous-sections précédentes ont mis en avant les travaux de la communauté pour avoir des motifs plus riches, permettant ainsi de mieux capturer les dimensions spatiales et temporelles. Les domaines de motifs étant de plus en plus complexes et le nombre de solutions potentielles de plus en plus important, beaucoup d'efforts ont été portés sur le développement d'algorithmes performants et la définition de contraintes permettant d'avoir des motifs plus pertinents et moins nombreux. Pour cela, une grande partie des travaux se sont appuyés sur des propositions faites pour des domaines de motifs plus généraux ou plus simples, avec une adaptation aux données spatio-temporelles souvent non triviale. Malgré toutes ces avancées, les verrous restent encore nombreux. Dans la suite de ce manuscrit, plusieurs travaux sont présentés afin de lever (en partie au moins) certains de ces verrous. Nous nous sommes plus particulièrement intéressés à l'intégration de la connaissance du domaine d'application, et à la définition de domaines de motifs (et d'algorithmes) visant à mieux prendre en compte les dimensions spatiales, temporelles et d'analyse de phénomènes complexes.

Intégration de la connaissance du domaine Au delà du problème de passage à l'échelle, l'une des principales difficultés lorsque l'on fait de l'extraction de motifs est la pertinence des motifs extraits. Comme nous avons pu le constater dans les sous-sections précédentes, beaucoup de travaux s'intéressent à la définition de contraintes permettant d'avoir des motifs intéressants, avec en parallèle la recherche de propriétés permettant d'optimiser l'exploration de l'espace de recherche. Généralement, ces contraintes sont totalement indépendantes de l'application étudiée. Il s'agit de contraintes liées directement au motif (p.ex. taille, durée ou structure) ou de contraintes liées à des mesures statistiques sur les motifs (p.ex. la fréquence ou l'indice de participation). Peu de travaux ont étudié l'intégration de contraintes directement liées à la connaissance du domaine tels que [BVd$^+$06, Ant08]. De plus, les contraintes proposées intègrent peu les dimensions spatiales et thématiques des données. Elles sont appliquées en post-traitement [BVd$^+$06] ou pendant l'extraction mais sans pouvoir être toujours utilisées pour élaguer l'espace de recherche faute de propriétés théoriques (comme la monotonie de la fréquence) [Ant08]. Par ailleurs, elles sont définies manuellement par les experts, ce qui nécessite un fort investissement de leur part.

Dans ce contexte, nous proposons dans le chapitre 1 de la partie III différents types de contraintes spatiales et thématiques pour faire de l'extraction de co-localisations guidée par le domaine. Toutes ces contraintes ont la propriété d'être anti-monotones et peuvent donc être utilisées pendant l'extraction pour améliorer l'efficacité des algorithmes. Nous décrivons aussi une approche permettant de dériver des contraintes à partir des modèles développés par les experts dans la littérature du domaine. Ces modèles (des fonctions mathématiques de plusieurs variables) représentent la connaissance des experts sur le phénomène étudié. Ils permettent donc de remplacer une grande partie des contraintes qui auraient été définies manuellement par ces derniers, voire de les compléter.

Extraction de motifs spatio-temporels plus riches Une autre difficulté importante est la définition d'un domaine de motifs suffisamment riche pour pouvoir capturer toutes les interactions possibles entre les dimensions d'analyse, tout en prenant en compte les spécificités des dimensions spatiales et temporelles. Au delà de la définition même du domaine de motifs, la principale difficulté est de mettre en place des algorithmes efficaces et complets pour extraire des motifs structurellement très complexes. Les sous-sections précédentes ont présenté un grand nombre de travaux ayant étudié cette problématique. Ces travaux se focalisent sur certains types de données en entrée et sur certains types d'interactions.

Par exemple, différents travaux ont recherché à extraire des évolutions fréquentes dans une base de données spatio-temporelles. Comme décrit dans la sous-section 2.2.1, les co-localisations ont été étendues de différentes façons afin de mettre en avant certaines évolutions et interactions dans l'espace et dans le temps. Toutefois, ces travaux font l'hypothèse que chaque objet spatial n'est décrit que par un seul attribut (son type). De même, les travaux sur les motifs séquentiels ont été étendus à ce contexte (p.ex. [TG01]). Ils permettent de considérer l'ensemble des attributs associés à chaque objet, mais ne considèrent que très partiellement les interactions spatiales. Face à ces limites, nous présentons un nouveau domaine de motifs dérivé des motifs séquentiels qui permet d'intégrer les informations des objets voisins (cf. chapitre 2 de la partie III). Ces motifs, appelés motifs spatio-séquentiels, sont extraits en utilisant une adaptation de l'algorithme $PrefixSpan$ [PHMA$^+$01]. Cette adaptation intègre de nouveaux types d'extensions liés à la relation spatiale et considère aussi le voisinage lors de la projection des données.

Les travaux précédents font l'hypothèse que les objets étudiés sont fixes dans l'espace (p.ex. des quartiers d'une ville). Toutefois, cette hypothèse peut ne pas s'appliquer dans certains cas d'étude (p.ex. le suivi de l'érosion des sols). En effet, les objets peuvent connaître des dynamiques complexes telles que des apparitions/disparitions ou des fusions/divisions. Certains travaux sur les co-localisations spatio-temporelles peuvent analyser ce type de données, mais toujours avec la contrainte d'avoir un attribut par objet. Face à ce problème, nous proposons dans le chapitre 3 de la partie III de modéliser dans un premier temps les données sous la forme d'un unique graphe orienté acyclique attribué. Beaucoup de travaux ont étudié l'extraction de motifs dans une collection de graphes (dans certains cas attribués), mais aucun n'a considéré un unique graphe attribué (cf. section 2.1.2). Ce type de données est plus complexe à analyser car les occurrences des motifs sont souvent entrelacées. Dans un second temps, nous proposons donc d'extraire des chemins fréquents dans ce type de graphe, ce qui n'a jamais été étudié jusqu'à présent. Même s'il s'agit structurellement de séquences d'*itemsets* sous contraintes, ces motifs sont difficiles à extraire en raison de leur entrelacement dans un unique et même graphe. Pour cela, nous avons une nouvelle fois adapté la stratégie de $PrefixSpan$ [PHMA$^+$01]. Néanmoins, contrairement au travail précédent, l'algorithme est profondément modifié tant au niveau des extensions qu'au niveau des projections effectuées. L'une des plus grandes difficultés est que l'extension d'une solution en cours de construction peut impliquer une modification de son préfixe, voire la génération d'un nouveau. Nous avons donc simultanément des extensions des préfixes et des suffixes. Cette particularité a nécessité le développement d'une structure de données optimisée spécifique pour garantir le passage à l'échelle.

La représentation de l'espace sous forme de graphe est assez classique dans la littérature (p.ex. graphe de voisinage). Elle permet de représenter l'ensemble des objets et leurs relations spatiales. Si chaque objet est associé à un ensemble d'attributs, on obtient un graphe attribué. L'évolution d'un tel graphe dans le temps est particulièrement intéressante car elle permet de représenter un grand nombre d'interactions et de dynamiques complexes. Comme nous l'avons vu précédemment, peu de travaux étudient ces graphes dynamiques attribués. En effet, ils sont parmi les structures les plus difficiles à analyser actuellement car ils combinent la complexité des

itemsets, des séquences et des graphes, le tout dans le cadre du graphe unique. Pour pouvoir traiter de tels graphes, les travaux actuels prennent des hypothèses fortes sur les données en entrée (p.ex. noeuds et arêtes fixes) et/ou se focalisent sur des domaines de motifs très spécifiques (p.ex. des co-évolutions). Dans ce contexte, nous proposons un algorithme en section 4 de la partie III permettant d'extraire des évolutions récurrentes de sous-graphes attribués. Plus précisément, il s'agit de séquences de sous-ensembles de noeuds attribués. La configuration exacte des arêtes n'est pas considérée afin d'avoir des motifs plus généraux, seule la connexité entre les noeuds est imposée. L'algorithme développé adopte une stratégie incrémentale totalement originale basée sur des intersections de graphes. Ces intersections ont certaines propriétés théoriques permettant de générer directement des morceaux de solutions. Il suffit ensuite de les associer temps après temps.

Restitution synthétique et intuitive des solutions Au delà de la phase d'extraction, la phase de restitution des résultats aux experts est également une étape clé du processus de découverte de connaissances. En effet, l'interprétation des solutions extraites par des experts de domaines totalement différents peut être difficile. Malgré les différentes contraintes intégrées, le nombre de motifs extraits peut aussi être très important, ce qui complique encore leur interprétation.

La visualisation des résultats issus de l'extraction est une problématique importante en fouille de données. En pratique, cette restitution se limite souvent à afficher la liste des motifs extraits avec les mesures d'intérêt associées (p.ex. la fréquence), ce qui peut être difficile à appréhender pour les experts. Plusieurs travaux se sont intéressés à cette problématique tels que [LIC08, BL10], notamment dans le cadre de données spatio-temporelles [AA99]. Une des principales approches proposées consiste à sélectionner un motif dans la liste des solutions, puis à afficher toutes ses occurrences sur une carte. Cette approche nécessite donc de sélectionner les motifs les uns après les autres, sans autre information que le motif lui-même. Dans le chapitre 1 des contributions, nous proposons une approche différente pour visualiser les co-localisations extraites. Elle s'appuie sur un *clustering* des occurrences des motifs pour résumer les principales localisations (et leur configuration), le tout dans une seule et même carte. Nous avons ainsi une visualisation globale et synthétique de l'ensemble des solutions, tout en fournissant des informations spatiales et thématiques complémentaires. Plus généralement, dans chacune des contributions présentées dans ce manuscrit, nous avons travaillé au développement d'outils et d'interfaces de visualisations qui soient adaptés aux pratiques et besoins des experts. Néanmoins, nous avons opté pour des approches de visualisation plus classiques dans ces autres travaux.

Les sous-sections précédentes présentent également un certain nombre de travaux visant à afficher moins de solutions aux utilisateurs. Il s'agit notamment de supprimer les informations redondantes en proposant des représentations condensées des solutions (sans perte d'information). Une grande partie de ces contributions s'articulent autour de la notion de fermeture. Toutefois, définir une telle représentation dans le contexte de la fouille d'un graphe unique est plus complexe. Le concept de fermeture d'un motif s'appuie sur une relation entre les motifs et les transactions qui les contiennent (correspondance de Galois). Or, cette relation n'existe plus dans le cadre du graphe unique. Dans nos travaux sur les graphes, nous introduisons donc un nouveau concept basé sur le même principe : la non-redondance. Cette notion est intégrée sous forme de contrainte dans les deux algorithmes proposés.

Dans le travail sur les motifs spatio-séquentiels (chapitre 2 de la partie III), nous présentons également une nouvelle contrainte permettant de filtrer en post-traitement des motifs contradictoires, ce qui permet d'avoir moins de solutions en sortie et des solutions plus pertinentes. Cette

notion est directement dérivée des travaux de [Azé03] réalisés dans le cadre de l'extraction de règles d'association. Le principe est de filtrer les séquences associées aux mêmes *itemsets* mais apparaissant dans un ordre différent.

Troisième partie

Contributions

1

Extraction de co-localisations guidée par le domaine

L'extraction de motifs spatiaux, et plus particulièrement de co-localisations, est une des problématiques importantes en fouille de données spatiales. Pour rappel, une co-localisation est un ensemble d'évènements ou d'objets (des caractéristiques booléennes) apparaissant fréquemment à proximité [SH01]. Différents algorithmes ont été développés pour extraire ces motifs. Toutefois, l'interprétation des résultats par les experts du domaine est difficile en raison du grand nombre de motifs non intéressants habituellement découverts.

A mon arrivée en 2008, je me suis intéressé à cette problématique dans le cadre de l'analyse de données environnementales. La Nouvelle-Calédonie est un point chaud de biodiversité, tant au niveau terrestre que marin. Pour autant, il s'agit d'un territoire où l'exploitation minière (notamment le nickel) a un poids économique important. Le suivi, la protection et la valorisation de cet environnement sont donc des axes majeurs pour les équipes de recherche locales. L'érosion des sols (naturelle et anthropique) est une problématique au centre de beaucoup de préoccupations et donc d'études. Les données collectées lors de certaines de ces études ont été le point de départ de notre travail. L'objectif était de les utiliser pour mettre en avant les facteurs en lien avec cette érosion (et les quantifier). Pour cela, nous sommes partis des travaux sur la fouille d'*itemsets* et de co-localisations, et avons travaillé à l'intégration de nouvelles contraintes permettant d'avoir des motifs plus pertinents, tout en améliorant le passage à l'échelle. Une fois les motifs extraits, se pose encore le problème de leur restitution aux experts. Dans un second temps, nous avons donc étudié la question de la visualisation des résultats.

Le tableau 1.1 présente les différents étudiants (stagiaires et doctorants) ayant travaillé sur cette problématique. Il présente aussi les projets de recherche et les collaborations associés à ce travail. Suite à ce tableau sont également listées les principales publications.

Ce chapitre est organisé de la manière suivante. La section 1.1 introduit le cadre théorique de ce travail. La section 1.2 présente une première contribution visant à intégrer des contraintes spatiales et thématiques lors de l'extraction de co-localisations. La section 1.3 décrit une deuxième contribution visant à tirer partie de la connaissance encodée dans des modèles experts. Une proposition permettant d'afficher sur une carte synthétique les co-localisations intéressantes est présentée en section 1.3. Pour finir, la section 1.5 discute des résultats obtenus sur des jeux de données liés à l'érosion des sols.

Master/Thèse
C. Grison (2009) stage Université Lyon 2 co-encadrement : N. Selmaoui-Folcher (UNC)
E. Desmier (2009-2010) année de césure INSA Toulouse co-encadrement N. Selmaoui-Folcher (UNC)
C. Paul-Hus (2011) stage Université de Sherbrooke co-encadrement N. Selmaoui-Folcher (UNC)
J. Sanhes (2011-2014) thèse dirigée par N. Selmaoui-Folcher (UNC) et J.-F. Boulicaut (INSA Lyon), co-encadrant : F. Flouvat

Projets
projet CNRT "Fonctionnement des petits bassins versants" (2010-2014)
ANR FOSTER (2011-2014) "FOuille de données Spatio-Temporelles : application à la compréhension et à la surveillance de l'ERosion"

Collaborations
INSA Lyon, CNRS, IRD

TABLE 1.1 – Synthèse des encadrements, des projets et des collaborations en lien avec l'extraction de co-localisations guidée par le domaine

Principales publications
Frédéric Flouvat, Nazha Selmaoui-Folcher, Dominique Gay, Isabelle Rouet and Chloé Grison. Constrained colocation mining : application to soil erosion characterization. In Proceedings of the ACM Symposium on Applied Computing (SAC'10), Data Mining Track, Sierre, Switzerland, pp.1055- 1060, March 22-26, 2010.
Nazha Selmaoui-Folcher, Frédéric Flouvat, Dominique Gay and Isabelle Rouet. Spatial pattern mining for soil erosion characterization. In the International Journal of Agricultural and Environmental Information Systems (IJAEIS), Special Issue on Environmental and agricultural data processing for water and territory management, IGI Global, Vol. 2, N 2, Jully 2011.
Elise Desmier, Frédéric Flouvat, Dominique Gay and Nazha Selmaoui-Folcher. A clustering-based visualization of colocation patterns. In Proceedings of the International Database Engineering and Applications Symposium (IDEAS'11), Lisbon, Portugal, September 2011.
Frédéric Flouvat, Jérémy Sanhes, Claude Pasquier, Nazha Selmaoui-Folcher and Jean-François Boulicaut. Improving pattern discovery relevancy by deriving constraints from expert models. In Proceedings of the 21st European Conference on Artificial Intelligence (ECAI'14), Prague, Czech Republic, August 18-22, 2014.
Frédéric Flouvat, Jean-François N'guyen Van Soc, Elise Desmier and Nazha Selmaoui-Folcher. Domain-driven co-location mining : Extraction, visualization and integration in a GIS. In GeoInformatica, Vol. 19, p.147-183, 2015.

TABLE 1.2 – Synthèse des publications en lien avec l'extraction de co-localisations guidée par le domaine

1.1 Cadre théorique

1.1.1 Domaine de motifs, contrainte et problématique

Cette sous-section rappelle le cadre introduit dans [SH01, HSX04, YS06] pour l'extraction de co-localisations. Soient $\mathcal{F}$ un ensemble de types d'évènements (ou d'objets) et $\mathcal{D}$ une base de données d'objets spatiaux. Chaque objet dans $\mathcal{D}$ correspond à un tuple $(location, feature)$, où $location$ correspond à une localisation et $feature \in \mathcal{F}$ est un type d'évènements ou d'objets. Pour simplifier les exemples, nous considèrerons des objets spatiaux représentés par des points dans un espace euclidien à deux dimensions. Par exemple, dans la figure 1.1, $\mathcal{F} = \{a, b, c, d, e\}$, $\mathcal{D} = \{a_1, c_2, b_3..., e_{12}\}$ avec $a_1 = ((x_1, y_1), a)$, $c_2 = ((x_2, y_2), c)$, etc.

Une **co-localisation** $X \subseteq \mathcal{F}$ est un sous-ensemble de types d'évènements tel que ces instances sont localisées dans le même voisinage. La relation de voisinage est une relation binaire $\mathcal{R}(o, o')$ entre deux objets spatiaux o et o'. En fonction des besoins, $\mathcal{R}$ peut être basée sur une distance seuil entre les deux objets, ou basée sur leur intersection, ou n'importe quelle relation spatiale réflexive et symétrique (p.ex. l'inclusion ne satisfait pas ces propriétés).

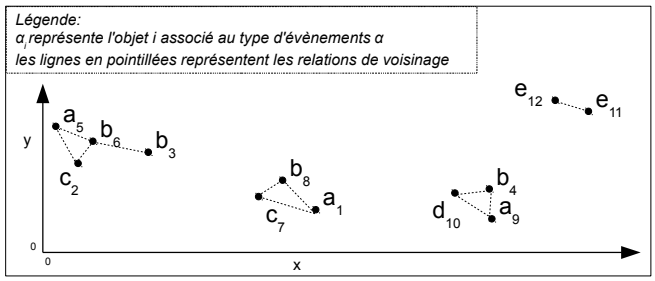


FIGURE 1.1 – Exemple de base de données spatiales

Une **instance de co-localisation** est un ensemble d'objets formant une clique (d'après $\mathcal{R}$). Par exemple, l'ensemble d'objets $\{a_9, b_4, d_{10}\}$ dans la figure 1.1 est une instance de la co-localisation $\{a, b, d\}$ par rapport à un seuil de distance euclidienne fixe (représenté en pointillés). A l'opposé, $\{a_1, b_4, c_7\}$ ou $\{a_1, b_4, d_{10}\}$ ne sont pas des instances de co-localisations de $\{a, b, d\}$. Le **tableau d'instances** de la co-localisation X, notée TI_X, est l'ensemble de toutes les instances de X.

Afin de filtrer des co-localisations intéressantes, les auteurs ont introduit une mesure d'intérêt, appelée **indice de participation** (*participation index*), qui correspond au minimum des probabilités d'apparition des évènements composant la co-localisation, i.e.

$$pi(\mathcal{D}, X) = \min_{\forall f \in X}(\ pr(\mathcal{D}, X, f) \ ), \text{ avec } pr(\mathcal{D}, X, f) = \frac{|\{o \in I \mid I \in TI_X \text{ et } o \text{ est de type } f\}|}{|TI_f|}$$

Par exemple, $pr(\mathcal{D}, \{a, b, c\}, a) = 2/3$ dans la figure 1.1, car il y a deux instances de type $\{a, b, c\}$ (i.e. $\{a_1, b_8, c_7\}$ et $\{a_5, b_6, c_2\}$) alors qu'il y a trois instances de type A au total (a_1, a_5 et a_9). De la même manière, $pr(\mathcal{D}, \{a, b, c\}, b) = 1/2$ and $pr(\mathcal{D}, \{a, b, c\}, c) = 1$. Nous obtenons donc $pi(\mathcal{D}, \{a, b, c\}) = 1/2$.

Au final, le problème à résoudre est le suivant : *Soient un ensemble de types d'évènements $\mathcal{F}$, une base de données spatiale $\mathcal{D}$, une relation de voisinage $\mathcal{R}$ et un seuil $\alpha \in [0, 1]$. L'objectif est de trouver l'ensemble des co-localisations dont l'indice de participation est supérieur à α, i.e.* $\{X \subseteq \mathcal{F} \mid pi(\mathcal{D}, X) \ \geq \ \alpha\}$.

1.1.2 Parallèle avec la fouille d'*itemsets*

[MTV97b] a introduit un cadre théorique généralisant le problème de la fouille d'*itemsets* fréquents. Ce cadre montre que les algorithmes de fouille d'*itemsets* peuvent être utilisés pour résoudre une grande variété d'autres problèmes telles que la découverte d'épisodes, la réécriture de requêtes, l'inférence de dépendances fonctionnelles ou d'inclusion. Ce cadre peut être résumé de la manière suivante : "Soient une base de données $\mathcal{D}$, un langage fini $\mathcal{L}$ pour exprimer des motifs ou définir des sous-groupes des données, et un prédicat anti-monotone (ou monotone) Q pour évaluer si un motif $\varphi \in \mathcal{L}$ est vrai ou "intéressant" dans $\mathcal{D}$, l'objectif est de trouver une théorie de $\mathcal{D}$ par rapport à $\mathcal{L}$ et Q, i.e. l'ensemble $Th(\mathcal{D}, \mathcal{Q}) = \{\varphi \in \mathcal{L} \mid Q(\mathcal{D}, \varphi) \text{ est vrai}\}$".

Le problème de découverte de co-localisations défini dans [SH01] est une autre application de cadre. Le langage est l'ensemble des combinaisons de types d'évènements et le prédicat est une conjonction de contraintes de voisinage et d'indice de participation. Une grande partie des algorithmes développés pour l'extraction d'*itemsets* fréquents peut donc être utilisée pour trouver les co-localisations. [SH01] ont ainsi pu appliquer la même stratégie que l'algorithme

Apriori [AS+94]. Ce cadre montre aussi qu'il est possible d'utiliser d'autres prédicats pour filtrer les co-localisations (en plus des contraintes d'indice de participation et de voisinage). La seule condition est d'avoir une conjonction de prédicats anti-monotone (ou monotone). Ainsi, il sera possible de les utiliser pour élaguer l'espace de recherche sans impacter l'efficacité de la stratégie d'exploration.

1.2 Intégration de contraintes spatiales et thématiques définies par les experts

1.2.1 Contraintes spatiales et thématiques

L'intégration de contraintes dans la fouille d'*itemsets* a largement été étudiée dans la littérature. Toutefois, ces contraintes (applicables aux co-localisations) ne considèrent pas les dimensions spatiale ou thématique des données géographiques, alors qu'il s'agit de concepts clés pour les experts. Face à ce manque, nous avons donc défini deux types de contraintes :

— des contraintes sur les types d'évènements/objets et leurs thèmes

— des contraintes spatiales sur les évènements/objets

Dans notre contexte, ces contraintes du domaine représentent des relations connues ou non intéressantes pour les experts. Ces contraintes peuvent être considérées comme des règles d'exclusion.

Les contraintes sur les types d'évènements et les thèmes Ce premier type de contraintes exclut des co-localisations avec des types d'évènements ou des thèmes particuliers. Le test dépend uniquement de l'analyse du motif (sans accès aux données). Nous avons défini six contraintes de ce type (cf. tableau 1.3). Toutes ces contraintes sont anti-monotones et peuvent être utilisées pour élaguer l'espace de recherche lors de l'extraction. De plus, d'autres contraintes du domaine peuvent être définies sur la base d'une conjonction ou d'une disjonction de celles-ci.

Contrainte	Type
$Q_{allFeatures}(X,F) = \neg(F \subseteq X)$	types d'évènements
$Q_{features}(X,F) = \neg(F \cap X \neq \emptyset)$	types d'évènements
$Q_{theme}(X,t) = \neg(X \cap t \neq \emptyset)$	thème
$Q_{intra}(X,t) = \neg(\lvert X \cap t \rvert \geq 2)$	intra-thème
$Q_{inter}(X,T) = \neg(\forall t \in T(X \cap t \neq \emptyset))$	inter-thème
$Q_{partInter}(X,T) = \neg(\exists t_1 \in T(t_1 \cap X \neq \emptyset) \wedge \exists t_2 \in T(t_2 \cap X \neq \emptyset))$	inter-thème partiel

TABLE 1.3 – Contraintes sur les types d'évènements et les thèmes Q_{Evt}

Les contraintes $Q_{allFeatures}(X,F)$ et $Q_{features}(X,F)$ permettent d'exclure des co-localisations contenant (tout ou en partie) des types d'évènements d'un ensemble F. Par exemple, si $F = \{"serpentinite", "harzburgite"\}$, alors toutes les co-localisations composées de $\{"serpentinite", "harzburgite"\}$ sont ignorées pendant la fouille.

Les autres contraintes sont des contraintes thématiques. Par exemple, la contrainte $Q_{theme}(X,t)$ exclut une co-localisation si elle est composée de types d'évènements du thème t. Par exemple, si t correspond au thème "végétation", la co-localisation $\{$*"savane", "sol nu", "serpentinite"*$\}$ ne sera pas étudiée. Les trois contraintes suivantes correspondent à des contraintes intra et inter-thèmes. La contrainte $Q_{intra}(X,t)$ exclut les co-localisations entre des types d'évènements du même thème t. Par exemple, l'expert peut ne pas être intéressé par les motifs mettant

en avant des corrélations entre *"serpentinite"* et *"harzburgite"* (différents types de sols) dans le thème *"lithologie"*. La contrainte $Q_{inter}(X, T)$ exclut les co-localisations entièrement liées à un ensemble de thèmes T. Par exemple, l'expert peut ne pas être intéressé par les co-localisations associées aux thèmes *"érosion"* et *"constructions humaines"*, et vouloir se focaliser uniquement sur l'érosion naturelle. La dernière contrainte est plus forte que $Q_{inter}(X, T)$, car elle exclut un motif s'il est associé à au moins deux thèmes de T (et non tous les thèmes).

Les contraintes spatiales sur les évènements Ce second type de contraintes exclut des objets spatiaux. L'idée est d'éviter (ou au contraire de centrer) l'analyse de corrélations dans des zones géographiques définies par l'utilisateur. Nous avons défini la contrainte $Q_{spatialAll}(I, shape, r)$ pour cela (cf. tableau 1.4). Cette contrainte permet de sélectionner uniquement les instances I d'une co-localisation si elles satisfont la relation spatiale r dans la zone *shape* (p.ex. représentée par un polygone). Grâce à ce prédicat, il est possible d'exprimer des contraintes tel que "étudier toutes les corrélations à proximité d'une zone minière".

Contrainte	Type
$Q_{spatialAll}(I, shape, r) \equiv (\forall o \in I(r(o, shape)))$	toute relation spatiale booléenne
$Q_{spatialAll}(I, shape, r) \vee CF$ avec CF une contrainte sur les types d'évènements et les thèmes du tableau 1.3	spatiale et thématique
$Q_{spatialAll}(I, shape, r) \wedge CF$ avec CF une contrainte sur les types d'évènements et les thèmes du tableau 1.3	spatiale et thématique

TABLE 1.4 – Contraintes spatiales Q_{Spa}

A noter que tous les objets dans I doivent satisfaire la relation spatiale. Par exemple, si r est la relation *"pas dans"*, tous les objets dans I ne doivent pas être dans la zone *shape*. Une instance dont seuls certains évènements ne se trouvent pas dans la zone n'est pas étudiée. En effet, une contrainte spatiale telle que "$\exists o \in I(r(o, shape))$" ne peut pas être utilisée car l'indice de participation de la co-localisation X pourrait être supérieur à celui de $Y \subseteq X$, et le prédicat ne serait donc pas anti-monotone (non directement exploitable lors de l'extraction).

Les deux dernières contraintes du tableau montrent que la contrainte spatiale peut être combinée avec des contraintes sur les types d'évènements et les thèmes. Ce type de contraintes peut être utilisé pour éviter l'analyse de corrélations spécifiques dans des zones spécifiques.

Contrairement aux contraintes sur les types d'évènements et les thèmes, les contraintes spatiales ne sont pas utilisées directement pour élaguer les co-localisations. Ces contraintes affectent le calcul de l'indice de participation en réduisant le nombre d'instances étudiées. Ainsi, elles ne sont pas intégrées dans le prédicat utilisé dans l'algorithme d'extraction, mais elles modifient le calcul du tableau d'instances effectué dans ce dernier. De plus, elles ne modifient pas l'anti-monotonie du prédicat. En effet, le nombre d'instances utilisées pour traiter l'indice de participation diminue toujours (non strictement) chaque fois que nous avons une conjonction ou une disjonction des contraintes spatiales précédentes.

1.2.2 Intégration dans un algorithme d'extraction de motifs

Le cadre théorique discuté en section 1.1 montre qu'il est possible d'intégrer directement ces contraintes spatiales et thématiques définies par les experts dans les algorithmes d'extraction d'*itemsets*. A titre d'exemple, cette section montre comment cette intégration se fait dans l'algorithme classique *Apriori* défini dans [AS+94], généralisé dans [MTV97b] et utilisé dans [SH01]

pour extraire les co-localisations. Un autre exemple d'intégration dans un algorithme d'extraction de motifs est présenté dans [FNVSDSF15].

L'algorithme 1 illustre l'intégration des contraintes spatiales et thématiques dans cet algorithme. Les contraintes spatiales sur les objets sont utilisées dans l'étape d'évaluation (lignes 4 à 8), c'est-à-dire lorsqu'on vérifie si une co-localisation est intéressante ou non par rapport au seuil d'indice de participation. Ces contraintes limitent le nombre d'objets étudiés lors de la génération du tableau d'instances de chaque co-localisation (ligne 5), et donc le nombre de jointures spatiales effectuées. Les contraintes sur les types d'évènements et les thèmes sont utilisées dans l'étape de génération (ligne 11), i.e. lors de la construction de nouvelles co-localisations candidates à partir de celles intéressantes trouvées dans la précédente itération. Ces contraintes suppriment, de l'ensemble des motifs candidats, les co-localisations ne satisfaisant pas les contraintes thématiques définies par l'expert.

Algorithm 1 Algorithme par niveau d'extraction de co-localisations basé sur des contraintes

Require: une base de données spatiales $\mathcal{D}$, un ensemble de type d'évènements $\mathcal{F}$, une relation spatiale $\mathcal{R}$, le seuil d'indice de participation α, les contraintes sur les évènement et les thèmes Q_{Evt}, et les contraintes spatiales Q_{Spa}

Ensure: toutes les co-localisations intéressantes vérifiant les contraintes, i.e. $Th(\mathcal{D}, Q_{Evt} \wedge Q_{Spa})$

1: $Cand_1 = \{f \in \mathcal{F} \mid Q_{Evt}(X) = vrai\}; k = 1$
2: **while** $Cand_k \neq \emptyset$ **do**
3: // ***Evaluation des co-localisations par rapport aux contraintes spatiales***
4: **for all** $X \in Cand_k$ **do**
5: $d' = \mathcal{D} \backslash \{o \in \mathcal{D} \mid$ o est un objet spatial d'une instance I de la co-localisation X par rapport à $\mathcal{R}$ et $Q_{Spa}(I) = faux\}$
6: **if** $pi(d', X) \geq \alpha$ **then**
7: $Th = Th \cup \{X\}$
8: **end if**
9: **end for**
10: // ***Génération des co-localisations en tenant compte des contraintes sur les types d'évènements et leur thème***
11: $Cand_{k+1} = \{X \subseteq \mathcal{F} \mid |X| = k+1 \wedge \forall Y \subset X,\ Y \in Th \wedge Q_{Evt}(X) = vrai\}$
12: $k = k + 1$
13: **end while**
14: **Return** Th

1.3 Intégration de contraintes du domaine dérivées de modèles des experts

Les contraintes précédentes permettent aux utilisateurs de cibler l'analyse et de rendre l'extraction plus performante. Elles représentent souvent une partie de la connaissance du domaine sur un phénomène donné. Toutefois, leur définition nécessite une forte implication des experts et plusieurs itérations du processus KDD. Afin de limiter cela, nous proposons de dériver une partie de ces contraintes de modèles mathématiques définis par les experts dans la littérature.

En effet, les experts de domaines scientifiques variés (e.g., des géologues, des physiciens ou des épidémiologistes) définissent et utilisent des modèles mathématiques pour représenter leur connaissance des phénomènes étudiés (principalement pour faire de la simulation). Par exemple, les experts en érosion des sols ont développé des modèles permettant d'estimer le risque d'érosion en fonction d'un ensemble de paramètres environnementaux [Mor01, Ath05]. De même les épidémiologistes ont défini des modèles afin d'estimer le nombre de personnes potentiellement infectées par une maladie [dCB+11]. Ces modèles experts présentent l'avantage de synthétiser

une partie de la connaissance du domaine dans un contexte donné. Il serait donc intéressant de les exploiter pour dériver des contraintes.

Nous nous intéressons plus particulièrement aux modèles mathématiques qui prennent la forme de fonctions à plusieurs variables (linéaires ou non), et prendrons pour exemple les modèles développés par les experts en érosion des sols.

1.3.1 Les modèles des experts

Le cas d'étude de l'érosion des sols

Les géologues et les géographes ont développé des modèles mathématiques visant à estimer le risque d'érosion d'un sol. Deux grandes classes de modèles peuvent être distinguées : les modèles empiriques et les modèles physiques.

Les modèles empiriques sont construits à partir de connaissances expertes et d'expérimentations. Le modèle USLE (*Universal Soil Loss Equation*) [WS78] et le modèle proposé dans [Ath05] en sont des exemples typiques (le premier est un modèle polynomial et le second est linéaire). Le modèle d'Atherton s'appuie sur deux indices : l'indice relatif de prédiction de l'érosion (REP) et le degré d'impact des infrastructures sur le bassin versant (WDI).

Paramètres	**classes**	**valeur**
Erodibilité du sol x_{erod}	alluvion	1
	sable, sol latéritique, ...	2
	marécage, limon nigrescent, ...	3
	sol latéritique ferrugineux,..	4
Occupation du sol x_{occup}	eau	0
	forêt dense, production de bois	1
	forêt clairsemée	2
	cocoteraie , zone non-forestière	3
	culture canne à sucre	4

Paramètres	**classes**	**valeur**
Pente (en %) x_{pente}	[0 , 3.5]	0.5
	[3.6 , 30]	1
	[31 , 50]	2
	[51 , 60]	3
	[60.1 , 100]	9
Intensité des pluies (en mm/an) x_{pluie}	[0 , 2000]	1
	[2001 , 3200]	2
	[3201, 10 000]	3
Saisonnalité des pluies (en mm) x_{saison}	[0 , 70]	1
	[71, 200]	2

$$REP(x_{pente}, x_{pluie}, x_{saison}, x_{erod}, x_{occup}) = x_{pente} + x_{pluie} + x_{saison} + x_{erod} + x_{occup}$$

$$REP = \begin{cases} [6, 9.5[\longmapsto \text{score } Faible \\ [9.5, 11[\longmapsto \text{score } Moyen \\ [11, 12[\longmapsto \text{score } Fort \end{cases}$$

FIGURE 1.2 – Calcul de l'indice REP (*Relative Erosion Prediction*) du modèle Atherton

Les modèles physiques sont des modèles quantitatifs fondés sur des propriétés physiques et calibrés à partir des données expérimentales. Par exemple, les modèles WEPP (*Water Erosion Prediction Project*) [LN89] et RMMF (*Revised Morgan-Morgan Finney*) [Mor01] sont basés sur plusieurs modèles physiques qui sont non linéaires et non polynomiaux. Le modèle RMMF divise le processus d'érosion en deux phases : détachement par gouttes de pluie (cf. exemple en Figure 1.3) et détachement par ruissellement. Chaque phase est liée à un sous-modèle physique. Les résultats des deux sous-modèles sont ensuite additionnés pour estimer la perte en sol annuelle.

Formalisation des modèles experts

Soit l'ensemble de variables/attributs du modèle expert, noté $Dim_{model} = \{x_1, x_2, ..., x_n\}$. L'ensemble des valeurs possibles de l'attribut x_j correspond au domaine de x_j, i.e. $dom(x_j)$. Un **modèle mathématique** est une fonction

$$g : dom(x_1) \times dom(x_2) \times ... \times dom(x_n) \to \mathbb{R},\ x = (x_1, x_2, \ldots, x_n) \mapsto g(x)$$

Paramètres	**Domaine de valeurs**
Indice de détachement du sol (en g/J) x_K	défini en fonction du type de sol
Précipitation annuelle (en mm) x_R	[0 , 12 000]
Proportion de pluie interceptée par la végétation x_A	[0 , 1]
Pourcentage de couverture de la canopée x_{CC}	[0 ,1]
Intensité de la pluie (en mm/h) x_I	{10, 25, 30} en fonction du climat de la zone d'étude
Hauteur de la végétation (en m) x_{PH}	[0 , 130]

$$g(x_K, x_R, x_A, x_{CC}, x_I, x_{PH}) = x_k \times [x_R \times x_A \times (1 - x_{CC}) \times (11.9 + 8.7\ log\ x_I) + (15.8 + x_{PH}{}^{0.5}) - 5.87] \times 10^{-3}$$

FIGURE 1.3 – Modélisation du détachement par gouttes de pluie dans le modèle RMMF

Par ailleurs, pour qu'un modèle puisse être exploité lors de la fouille de données, la base de données doit contenir au moins un attribut du modèle expert, i.e. $Dim_{\mathcal{D}} \cap Dim_{model} \neq \emptyset$.

1.3.2 Des motifs aux modèles

Différents types de contraintes peuvent être dérivés des modèles en fonction des données, du type de modèles utilisés, mais aussi du problème étudié. Dans le cadre de cette sous-section, nous nous focalisons plus particulièrement sur une contrainte relativement similaire à une contrainte de fréquence minimale, même si ses propriétés sont très différentes. Cette contrainte vise à filtrer les motifs X tel que $g(X)$ est supérieure ou égale à un seuil, i.e.

$$Q_{g\geq}(X) \equiv g(X) \geq minf$$

Par exemple, si g estime la perte en sol (en kg/m^2 par an), cette contrainte permettra de ne conserver que les motifs susceptibles de représenter une perte en sol (et donc une érosion) supérieure à une certaine quantité. Dans cette application, on peut distinguer deux cas en fonction de la disponibilité de données "terrain" sur le phénomène modélisé (la perte en sol).

En l'absence de "vérité terrain", cette contrainte permettra de mettre en avant si de telles pertes risquent d'être fréquentes dans la zone d'étude et dans quelles situations. Elle peut également permettre d'identifier à quels autres facteurs, non couverts par le modèle, ces pertes en sols sont fréquemment liées dans les données.

En présence de "vérité terrain", cette contrainte permettra de comparer la prévision du modèle des experts à la réalité des données collectées. Les motifs confirmant le modèle sont intéressants car ils sont doublement validés par la vérité terrain et la connaissance du domaine (i.e., le modèle expert). De plus, les évènements additionnels dans le motif peuvent compléter les explications du modèle. Les motifs contredisant le modèle des experts sont tout aussi intéressants car ils montrent certaines spécificités non prises en compte dans les modèles experts utilisés, mettant donc en avant des possibilités d'améliorations.

Valeur d'un itemset X par un modèle g

La contrainte précédente nécessite de pouvoir calculer la valeur du motif par le modèle, i.e. $g(X)$. Par exemple, en considérant $g(x_1, x_2, x_3) = \sqrt{x_1} - cos(x_2)/2 \times log(x_3)$, si le motif X est {"$x_1 \in [3, 5]$", "$x_2 = 3$", "$x_3 = A$"}, quelle est la valeur de $g(X)$? Autrement dit, quelle est la prévision du modèle g pour les valeurs $x_1 \in [3, 5]$, $x_2 = 3$, et $x_3 = A$?

Pour une co-localisation comme {"x_1=1", "x_2 = 3", "x_3=A", *"mine"*}, il suffit de calculer $g(1, 3, 10)$, en supposant que "$x_3 = A$" est associé à la valeur 10 par les experts. Ce cas est simple

car toutes les variables du modèle sont exprimées dans le motif. De plus, elles sont associées à une unique valeur dans chacun des évènements. A noter que "*mine*" n'a pas besoin d'être considéré pour le calcul de g car il n'est pas pris en compte par le modèle. Cet élément est néanmoins intéressant car il apporte une information supplémentaire par rapport à la connaissance du modèle. Néanmoins, dans un cas plus général, cette opération soulève deux problèmes.

Premièrement, certaines variables du modèle peuvent ne pas être présentes dans les données ($Dim_{model} \not\subseteq Dim_{\mathcal{D}}$). De même, d'autres peuvent être absentes du motif car il n'implique qu'un sous-ensemble des valeurs apparaissant dans les données. Considérons la co-localisation X'= *{"x_1=1", "x_3=A", "mine"}*. Elle ne couvre pas toutes les variables du modèle (x_2 n'est pas exprimée). Il est tout de même possible de borner $g(X')$ en considérant les valeurs de x_2 pour lesquelles g est maximale/minimale. Dans notre exemple, si $x_2 = \pi$ ou 3π, alors $g(1,x_2,10)$=1.5. Cette valeur de g est la plus grande valeur possible étant donné les valeurs de x_1 et x_3. A l'opposé, si $x_2 = 0$, 2π ou 4π, alors $g(1,x_2,10)$=0.5. Cette valeur de g est la plus petite valeur possible étant donné les valeurs de x_1 et x_3. On en déduit que $0.5 \leq g(X') \leq 1.5$, même si x_2 n'est pas dans X'.

Deuxièmement, il est courant d'avoir des motifs représentant des mélanges d'intervalles, de valeurs numériques et de valeurs nominales. Considérons par exemple la co-localisation $X'' = \{\text{“}x_1 = 4\text{”}, \text{“}x_2 \in [0, 2\pi[\text{”}, \text{“}x_3 = A\text{”}\}$. Pour calculer $g(X'')$, il est possible d'appliquer une approche similaire à celle utilisée précédemment en bornant $g(X'')$ par rapport à $x_2 \in [0, 2\pi[$. En étudiant la fonction cosinus sur $[0, 2\pi[$, on sait que $g(X'')$ est maximale lorsque $x_2 = \pi$ (dans ce cas, $g(4, \pi, 10) = 2.5$) et minimale lorsque $x_2 = 0$ (dans ce cas, $g(4, 0, 10) = 1.5$). Nous pouvons donc en déduire que $1.5 \leq g(X'') \leq 2.5$.

Plus formellement, nous avons donc pour tout $i_j \in X$ représentant un intervalle $[inf_j, sup_j]$ d'une variable x_j du modèle g :

$$\min_{\forall i_j \in [inf_j, sup_j]} g(i_1, ..., i_j, ...i_n) \leq g(X) \leq \max_{\forall i_j \in [inf_j, sup_j]} g(i_1, ..., i_j, ...i_n)$$

Maintenant, il est important d'étudier les propriétés théoriques de ces contraintes afin de pouvoir les intégrer efficacement dans les algorithmes d'extraction.

Propriétés théoriques des modèles par rapport aux co-localisations

Trois propriétés théoriques ont notamment été identifiées. Elles permettent d'élaguer des motifs sans avoir à les générer ou les tester. De plus, leur coût de calcul est négligeable par rapport au gain apporté. Les deux premières propriétés permettent d'élaguer les sur-ensembles d'un motif sous certaines conditions liées au modèle étudié. La dernière propriété prend en compte la (dé-)croissance de la dérivée partielle de g par rapport à certains attributs pour élaguer des motifs partageant des attributs en commun.

Liens entre un motif et ses sur-ensembles Soit deux co-localisations $X, Y \subseteq \mathcal{F}$ tel que $X \subset Y$. Si X et Y ont les mêmes variables du modèle g, alors ils auront les mêmes valeurs pour ces variables, et par conséquent $g(Y) = g(X)$. En fait, Y ne se différencie de X que par des types d'évènements non pris en compte dans le modèle, ce qui n'influence pas le calcul de $g(Y)$. Dans ce cas, si $g(X) < minf$, alors $g(Y) < minf$.

Le cas est plus complexe lorsque des variables de g ne sont pas exprimées dans X mais le sont dans Y (seule autre possibilité si l'on conserve l'hypothèse $X \subset Y$). Prenons l'exemple du motif $X = \{\text{“}x_2 \in [0, 2\pi[\text{”}, \text{“}x_3 = A\text{”}\}$ où la variable x_1 n'est pas exprimée. Nous avons $0.5 = g(1, 0, 10) \leq g(X) \leq g(16, \pi, 10) = 4.5$, car $dom(x_1) = [1, 16]$. Considérons maintenant

un sur-ensemble $Y_1 = \{\text{“}x_1 = 16\text{”}, \text{“}x_2 \in [0, 2\pi[\text{”}, \text{“}x_3 = A\text{”}\}$, avec $g(Y_1) = [3.5, 4.5]$. Si le seuil minimum pour g est 2, $g(X) < minf$, pourtant $g(Y_1) > minf$. Par contre, si le seuil minimum est 5, on est sûr que tous les sur-ensembles de X vérifient $g(X) < 5$. On peut donc directement les éliminer. Cette propriété est appelée "cohérence des bornes" d'une contrainte (*bounds consistency*) dans la littérature.

Motifs partageant les mêmes attributs L'étude détaillée de g permet d'exposer d'autres propriétés entre les motifs. Toutefois, cette étude peut être complexe de par la nature des fonctions considérées (des fonctions à plusieurs variables non nécessairement linéaires). Il est difficile d'étudier globalement la monotonie d'une fonction à plusieurs variables. Notre solution consiste à analyser la fonction par rapport à une variable à la fois (les autres étant considérées comme des constantes). Cela revient à étudier les dérivées partielles de g pour chaque attribut et à identifier les intervalles dans lesquels la fonction est monotone. Dans chacun de ces intervalles (pour chacune des variables), il est possible de dériver des propriétés permettant d'élaguer l'espace de recherche.

Prenons par exemple les co-localisations X={"x_1=4", "$x_2 \in [\pi/2, \pi[$", "x_3=A"} et Y={ "x_1=4", “$x_2 \in [0, \pi/2[$”, "x_3 =A"}. L'analyse de la fonction g par rapport à x_2 montre qu'elle est strictement croissante sur $[0, \pi]$ (i.e., $\frac{\partial f}{\partial x_2} > 0$ sur $[0, \pi]$). Puisque X est plus grand que Y par rapport à x_2 , nous avons $g(Y) <g(X)$. En effet, g(X) = [2, 2.5[et g(Y) = [1.5, 2[. Par conséquent, si $g(X) < minf$, alors $g(Y) < minf$ (même si $X \not\subset Y$). De même, considérons le motif $Y'' = \{\text{“}x_1 = 1\text{”}, \text{“}x_2 \in [0, \pi/2[\text{”}, \text{“}x_3 = A\text{”}\}$. Nous avons $g(Y'') < g(X)$, car $\frac{\partial f}{\partial x_1} > 0$ sur $dom(x_1)$. Ainsi, si $g(X) < minf$, alors Y'' peut aussi être directement élagué.

Notons que l'impact de cette propriété dépend de la discrétisation. En effet, il n'aurait pas été possible de déduire cela si nous avions eu les motifs $X = \{\text{“}x_1 = 4\text{”}, \text{“}x_2 \in [\pi, 2\pi[\text{”}, \text{“}x_3 = A\text{”}\}$ et $Y = \{\text{“}x_1 = 4\text{”}, \text{“}x_2 \in [0, \pi[\text{”}, \text{“}x_3 = A\text{”}\}$, car g est croissante par rapport à x_2 sur $[0, \pi]$ et décroissante sur $[\pi, 2\pi]$. Soit une variable x_j dont le domaine est découpé en intervalles dans lesquels g est monotone. Plus le nombre de type d'évènements appartenant à chacun de ces intervalles est important, plus cette propriété est efficace.

Intégrer les modèles des experts dans l'extraction de motifs

La contrainte proposée est relativement simple à intégrer dans les algorithmes d'extraction de motifs, car elle possède des propriétés similaires à celles utilisées classiquement pour extraire des *itemsets* (p.ex. la contrainte de fréquence minimale). Elle impacte uniquement la génération des motifs candidats. Dans l'algorithme 1, elle est intégrée à la ligne 11 à la place, ou en conjonction, de la contrainte Q_{Evt}. L'intérêt d'utiliser cette contrainte de seuil basée sur des modèles experts lors de l'extraction (et non lors de l'étape de post-traitement) est d'élaguer rapidement les modèles inintéressants, ce qui améliore les performances et le passage à l'échelle.

Cette approche est totalement générique. La plupart des algorithmes d'extraction (p.ex. *Apriori*[AS$^+$94], *Eclat* [ZPO$^+$97], *FP-growth* [HPY00]) peuvent l'intégrer. Toutefois, selon la stratégie de l'algorithme, il peut s'avérer difficile d'exploiter certaines de ces propriétés pour élaguer l'espace de recherche. Par exemple, il est difficile de tirer partie de la dernière propriété dans des algorithmes dont la stratégie de génération des candidats est basée sur la fermeture (p.ex. *LCM*), alors que cela est plus facile pour les algorithmes tels qu'*Apriori*, *FP-growth* ou *Eclat* (puisque les motifs sont étendus par un élément à la fois et non plusieurs).

1.4 Visualisation cartographique des co-localisations

Un problème important en fouille de données est la présentation et la visualisation des résultats issus de l'extraction des données. Classiquement, les solutions sont retournées sous un format textuel, ce qui ne permet pas d'appréhender l'information spatiale des objets sous-jacents. La figure 1.4 illustre ce problème. Les objets à gauche sont utilisés pour extraire une liste de co-localisations intéressantes. Toutefois, cette liste seule contient une information spatiale très limitée. Elle permet seulement savoir que certains types d'évènements sont "souvent" proches. Elle ne permet pas de savoir où et comment ces évènements sont spatialement distribués. Au mieux, certaines approches retournent un rapport textuel avec la liste de toutes les solutions et affiche sur une carte les instances d'une seule co-localisation sélectionnée [AA99]. Cette approche permet d'avoir des informations détaillées sur les événements de la co-localisation sélectionnée, mais elle ne permet pas d'avoir une vue globale de toutes les co-localisations en même temps.

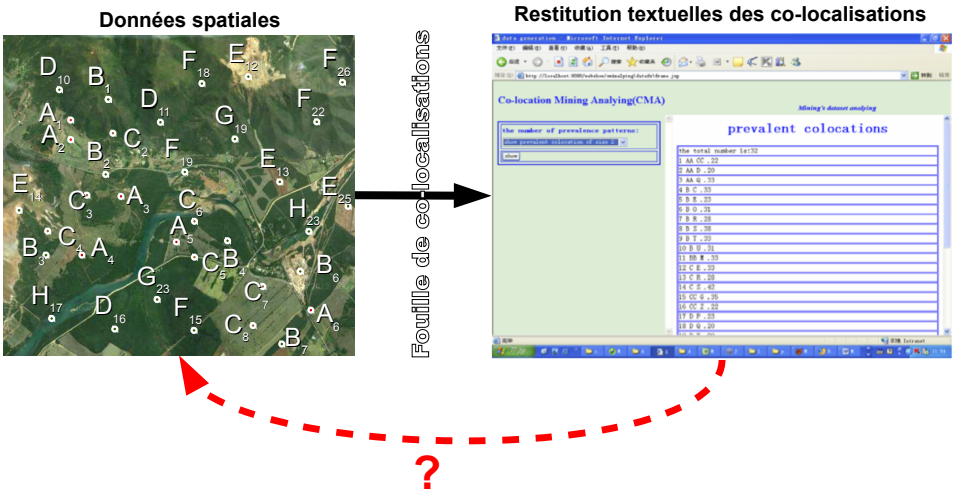


FIGURE 1.4 – Problème de la visualisation des co-localisations

Face à ce problème, nous proposons une nouvelle approche pour visualiser sur une carte l'ensemble des co-localisations extraites. Étant donné que chaque co-localisation intéressante peut comporter un grand nombre d'instances, notre idée est de les résumer en utilisant une nouvelle approche de *clustering* et de les intégrer dans une couche d'un SIG. La couche de co-localisations résultante indiquera aux experts où et comment chaque co-localisation est généralement située, donnant ainsi une vue globale de la distribution spatiale des solutions.

1.4.1 Comment représenter visuellement une co-localisation ?

Comme le montre la figure 1.5, il est naturel de représenter chaque co-localisation par une clique étiquetée, où chaque sommet représente un type d'évènements et chaque arête représente la relation de voisinage. La couleur des arêtes met en avant l'importance d'une co-localisation par rapport à la mesure d'intérêt (plus la couleur est vive, plus l'intérêt est élevé). La couleur des noeuds indique le thème associé au type d'évènements affiché (p.ex. vert pour le thème "Végétation").

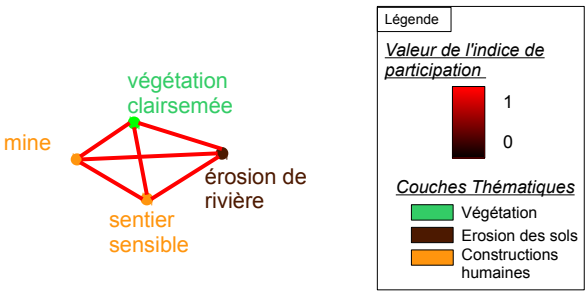


FIGURE 1.5 – Représentation sous forme de clique de la co-localisation *{"mine","végétation clairsemée","sentier sensible", "érosion de rivière"}* avec *pi({mining zone, "végétation clairsemée", "sentier sensible", "érosion de rivière"}) = 0.8*

1.4.2 Comment positionner une co-localisation sur une carte ?

Une co-localisation ne donne que peu d'informations spatiales. Par exemple, si une co-localisation $\{a, b, c\}$ est intéressante, alors l'expert sait uniquement que les types d'évènements (ou d'objets) a, b et c sont souvent proches les uns des autres, mais il ne sait pas où et comment. L'information spatiale d'une co-localisation est principalement portée par ses instances. Or, le nombre d'instances d'une même co-localisation peut être très important et leur distribution spatiale peut être hétérogène. Nous devons donc identifier certaines localisations typiques, i.e. grouper des instances en fonction de leur position spatiale. Pour cela, nous effectuons un *clustering* spatial.

Ce *clustering* peut être effectué en utilisant n'importe quel algorithme existant tel que *K-means* [Llo82] ou *DBSCAN* [EKSX96], directement pendant l'extraction. Toutefois, l'exécution d'un nouveau *clustering* pour chaque co-localisation prend beaucoup du temps. Nous pouvons optimiser ce traitement en considérant que toutes les co-localisations sont construites à partir du même ensemble de types d'évènements. Pour cela, nous proposons une heuristique en deux étapes, intégrée à l'algorithme d'exploration, basée sur :

— un *clustering* spatial des instances de chaque type d'évènements, effectué une seule fois au début de l'algorithme d'extraction.

— un *clustering* spatial des instances de chaque co-localisation basé sur les *clusters* précédents, en utilisant une approche "fusionner-diviser".

La première étape résume la localisation où chaque type d'évènements se produit. Pour cela, elle utilise l'algorithme *X-means* [PM00]. La sous-figure “clusters prétraités” de la figure 1.6 illustre ce "pré-*clustering*" dans lequel des instances de chaque type d'évènements (p.ex. a, b et c) sont partitionnées indépendamment par l'algorithme *X-means*.

La deuxième étape utilise ces *clusters* prétraités comme point de départ pour regrouper les instances de chaque co-localisation intéressante (cf. sous-figure “approche diviser-fuisionner” de la figure 1.6). Chaque tableau d'instances est traité en utilisant une approche de type "fusionner-diviser". Le principe est de partitionner ("diviser") les instances en fonction du type d'évènements f ayant le plus grand nombre de *clusters*. Cependant, en utilisant cette méthode, nous pouvons avoir des *clusters* conflictuels, i.e. deux *clusters* différents partageant des objets communs. La figure 1.6 illustre ce problème pour la co-localisation $\{a, b, c\}$. Si nous divisons en fonction des *clusters* de c présentés dans la sous-figure *3.* de la figure 1.6, nous avons $\{a_2, b_2, c_2\}$ et $\{a_3, b_2, c_3\}$ (deux instances de la co-localisation $\{a, b, c\}$) dans deux *clusters* différents (cf. sous-tableau *i.*

dans la figure 1.6). Cependant, ces deux *clusters* ont en commun l'objet b_2, ce qui signifie que ces deux groupes d'instances ne sont pas si éloignés l'un de l'autre (cf. sous-tableau *ii.*) dans la figure 1.6). Ils sont donc fusionnés (cf. sous-tableaux *iii.* et *iv.*). Ces opérations se répètent jusqu'à ce que le type d'évènements ayant le plus grand nombre de *clusters* ne change plus. Dans l'exemple de la figure 1.6, cette approche aboutit à la fusion des deux premiers *clusters* de c.

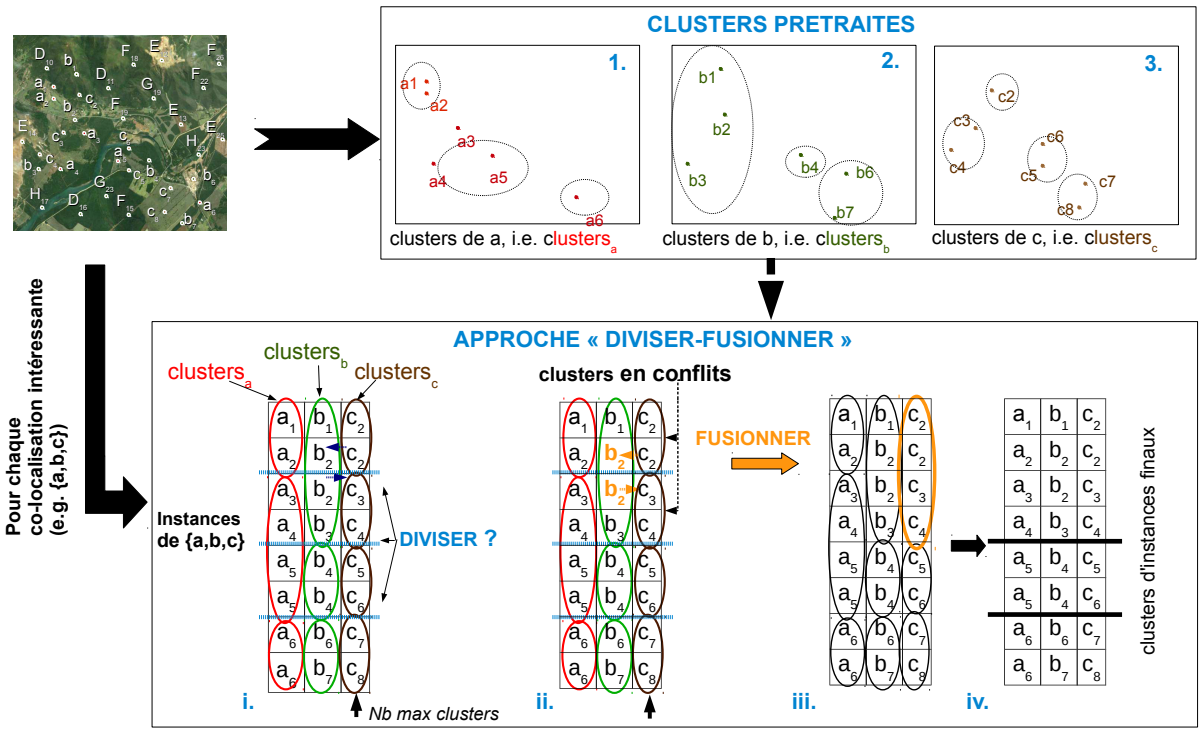


FIGURE 1.6 – *Clustering* des instances de la co-localisation $\{a, b, c\}$

Une fois ce *clustering* effectué, il suffit d'associer à chaque emplacement (i.e. chaque *cluster*) une clique et de positionner les sommets de cette clique en fonction des coordonnées spatiales moyennes des objets du *cluster*. Par exemple, dans la figure 1.7, le premier *cluster* est composé de quatre instances de la co-localisation $\{a, b, c\}$. Ces quatre instances impliquent quatre objets de type a ($\{a_1, a_2, a_3, a_4\}$), trois objets de type b ($\{b_1, b_2, b_3\}$) et trois objets de type c ($\{c_2, c_3, c_4\}$). Pour représenter cet emplacement typique de co-localisation $\{a, b, c\}$, nous représentons sur la carte une clique ayant trois sommets (un avec l'étiquette a, un avec l'étiquette b et un autre avec l'étiquette c). Chaque sommet est le centroïde des objets associés au type correspondant (p.ex., le sommet portant l'étiquette a est le centroïde des objets $\{a_1, a_2, a_3, a_4\}$). Cette approche est appliquée aux trois *clusters* de $\{a, b, c\}$, résultant en trois cliques dans la carte finale.

Le principal intérêt de cette approche est de visualiser plus précisément où et comment les co-localisations intéressantes sont généralement situées. Ainsi, elle fournit des informations supplémentaires aux experts par rapport aux solutions existantes. Par exemple, la figure 1.7 montre que la co-localisation $\{a, b, c\}$ est généralement située au nord-ouest, au centre et au sud-est de la carte. Cette approche a l'avantage de fournir aux experts une image globale de la distribution spatiale de toutes les co-localisations. De plus, elle fournit aussi des informations sur la manière dont les évènements d'une co-localisation sont les uns par rapport aux autres.

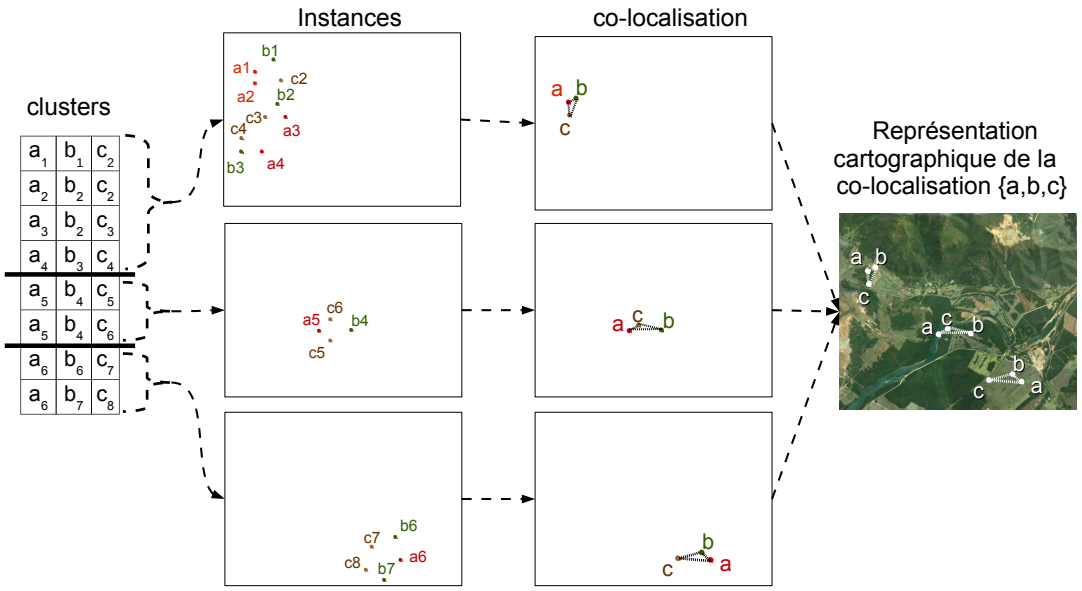


FIGURE 1.7 – Visualisation de la co-localisation $\{a, b, c\}$ sur une carte

1.5 Expérimentations et application à l'étude de l'érosion des sols

1.5.1 Prototype

Les propositions présentées précédemment ont été intégrées dans un prototype couplant une base de données *PostGIS*, la librairie d'extraction de motifs *iZi* [FDMP09], et l'outil de visualisation *Quantum GIS*. La figure 1.8 décrit l'architecture de ce prototype.

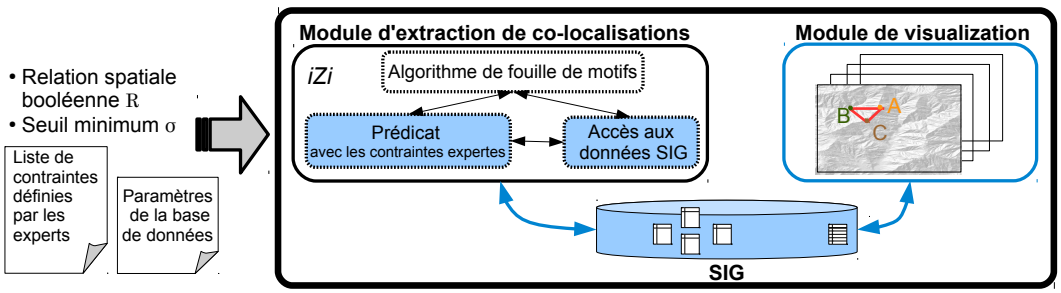


FIGURE 1.8 – Architecture du prototype

1.5.2 Protocole expérimental

Nous avons utilisé notre approche pour étudier l'érosion des sols dans deux zones impactées par l'érosion naturelle et anthropique (érosion liée aux activités minières). La première zone, dite de la "*Ouinné*", contient 68 types d'objets (liés à l'érosion, la couverture végétale, la géologie, l'activité minière et le réseau routier) et 3943 objets spatiaux. La seconde zone, dite de "*Kwe Binyi*", contient 71 types d'objets et 7306 objets spatiaux. Deux relations de voisinage ont

été considérées : l'intersection (fonction $St\_Intersects$ de $PostGis$) et la distance euclidienne (fonction $St\_Dwithin$).

Tout d'abord, un spécialiste de l'érosion des sols a analysé les co-localisations extraites par l'approche de [SH01] sur les jeux de données (sans intégrer aucune contrainte du domaine). L'interprétation des résultats a été faite à partir de l'affichage cartographique proposé dans la section précédente. Ensuite, l'expert à utilisé ces résultats pour définir des contraintes et éliminer les relations non intéressantes (car connues ou non pertinentes). Ces contraintes spatiales et thématiques ont été intégrées dans l'algorithme de [SH01] et étudiées en fonction des deux relations de voisinage. Pour finir, nous avons considéré une conjonction de contraintes basée sur la surface impactée par le motif et le modèle d'Atherton [Ath05] développé par des experts pour prédire le risque d'érosion. Afin de montrer la généricité de notre approche, nous avons intégré ces contraintes dans l'algorithme d'extraction de motifs *Close-By-One* [KO02].

Après cette analyse qualitative, nous avons étudié les performances (temps d'exécution et nombre de motifs) des algorithmes avec et sans les contraintes du domaine, ainsi que l'efficacité de notre approche de visualisation.

1.5.3 Analyse qualitative des motifs

Impact des contraintes définies par les experts

La co-localisation {*"cours d'eau secondaire", "sols nus sur substrat ultramafique"*} est un exemple de nouveau motif extrait grâce à ces contraintes. Il a été extrait avec la fonction $St\_Intersects$ et un seuil d'indice de participation de 0.5. Cette tendance est intéressante car ces sols (*"sols nus sur substrat ultramafique"*) sont souvent liés aux activités minières. Lorsque ces sols miniers sont dépourvus de végétation, l'érosion peut être très importante. Dans ce cas, les cours d'eau traversant ces sols peuvent être pollués.

L'expert a également découvert de nouveaux motifs intéressants avec $St\_Dwithin$ à 200 mètres et un seuil d'indice de participation de 0.5. Par exemple, la co-localisation {*"zone dégradée par les activités minières", "maquis lino-herbacé"*}, i.e. les zones dégradées par les activités minières sont souvent situées à proximité de zones couvertes de petits arbustes. La figure 1.9 présente ce motif ainsi que les données sous-jacentes. Les zones couvertes d'arbustes sont en jaune et les zones érodées sont entourées par un polygone orange. Ce motif est associé à deux cliques sur la carte (composées d'un point vert et d'un autre orange). La première clique représente les occurrences de la co-localisation situées en haut à gauche de la zone d'étude, alors que la deuxième représente les occurrences situées dans zone centrale. Ce type de végétation est intéressant en raison de son pourcentage élevé de plantes endémiques. De plus, il est particulièrement adapté aux sols miniers. Cette végétation est essentielle à la re-végétalisation et à la restauration de ces zones dégradées par les activités minières.

A noter que les motifs non intéressants extraits ont été à leur tour utilisés pour définir de nouvelles contraintes, affiner l'analyse et étudier des seuils de fréquence minimale plus faibles.

Impact des contraintes dérivées des modèles des experts

Plusieurs motifs intéressants pour les experts ont aussi pu être mis en avant grâce à la contrainte basée sur le modèle d'Atherton [Ath05]. D'après ce modèle, le risque le plus faible correspond à la valeur 1.5 et la plus élevée est 17. Ainsi, un seuil minimum de 3, i.e. $minf = 3$, est un seuil faible permettant d'ignorer lors de l'analyse les motifs non liés à un risque d'érosion. A l'opposé, un seuil minimum de 15, i.e. $minf = 15$, est un seuil élevé permettant de se focaliser sur les motifs liés à un fort risque d'érosion.

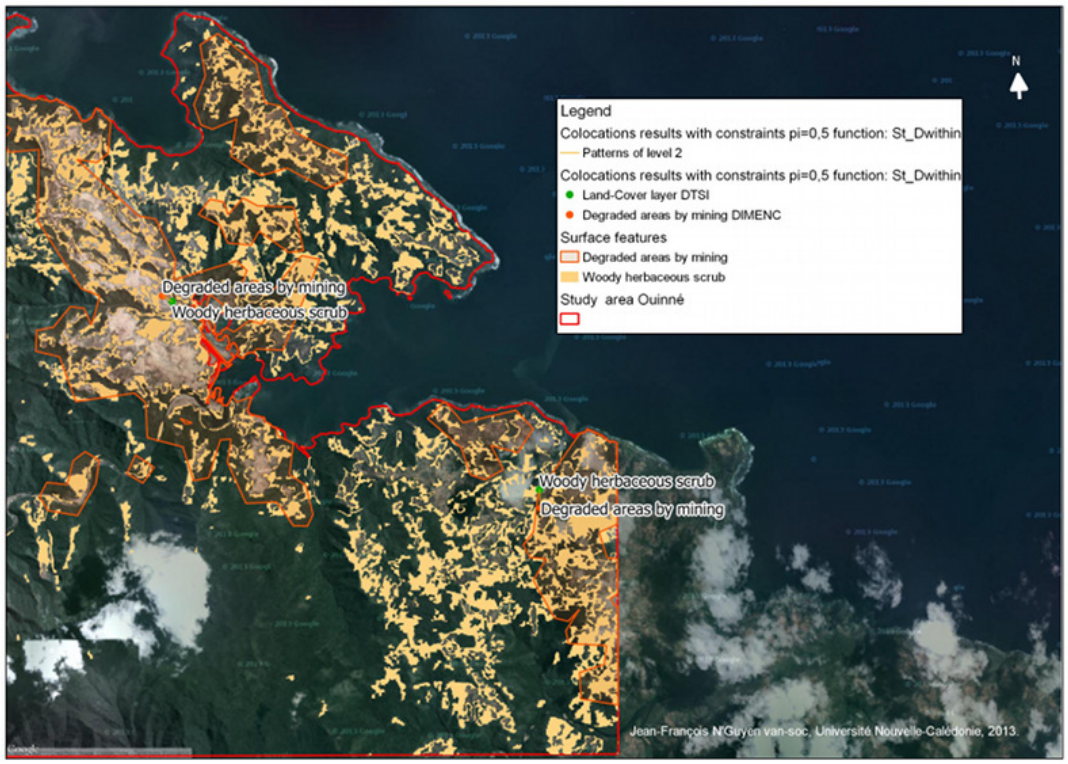


FIGURE 1.9 – Visualisation de la co-localisation {*"zone dégradée par les activités minières", "maquis lino-herbacé"*}

Par exemple, un motif couvrant 1% de la zone d'étude et lié à une forte érosion ($minf = 15$) a été extrait. Cette co-localisation est *{"serpentinite", "sol ultramafique sur substrat volcano-sedimentaire", "pente=[61,100]", "indice de végétation=(-0.071,0.115]"}*. Cette tendance montre qu'une partie de la zone étudiée est associée à un risque élevé d'érosion des sols. Ces zones à haut risque sont caractérisées par des sols avec de la serpentinite recouverts de substrat volcano-sédimentaire et présentant une pente importante. L'indice de végétation NDVI calculé à partir des informations radiométriques de l'image satellitaire confirme cela. La valeur de cet indice est faible, ce qui est typique d'une végétation clairsemée. Les attributs radiométriques ont un autre intérêt. Ces valeurs peuvent être utilisées sur d'autres images satellites pour identifier les zones à haut risque, même si nous n'avons ni la géologie ni la couverture terrestre (i.e. les données d'entrée du modèle) sur ces images.

Un autre exemple de motif est *{"latérites épaisses sur les péridotites", "maquis lino-herbacé", "pente=[3.6 ;30] "}*. Ce motif est associé à un risque modéré d'érosion des sols d'après le modèle des experts utilisé ($minf = 6$). Sa fréquence montre que 4-5 % de la région est caractérisée par un tel risque d'érosion.

1.5.4 Analyse quantitative

Impact des contraintes définies par les experts

Les résultats obtenus sur les deux jeux de données confirment l'impact des contraintes sur le temps d'exécution et le nombre de motifs extraits [1]. L'extraction des motifs est beaucoup plus efficace avec les contraintes définies par les experts, et le nombre de motifs est beaucoup moins important. Cette différence est plus importante lorsque la relation de voisinage est moins stricte (p.ex., *St_dwithin* à 200m), car davantage de co-localisations candidates sont générées et testées. La figure 1.10 illustre cela sur les données *Ouinné* et la relation de voisinage *St_dwithin* à 200m.

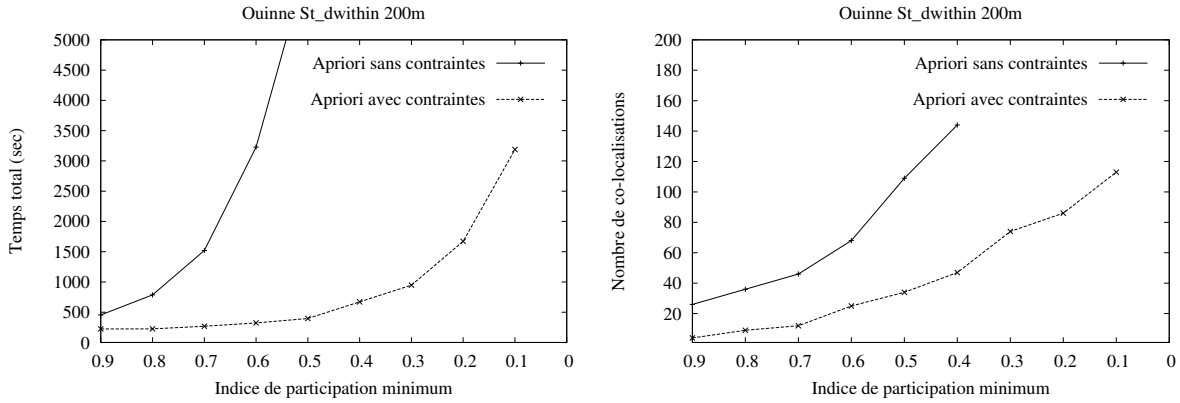


FIGURE 1.10 – Exemple de temps d'exécution et de nombre de co-localisations extraites avec et sans des contraintes définies par les experts (*Ouinné*)

Impact des contraintes dérivées des modèles des experts

Comme attendu, la contrainte basée sur le modèle des experts permet de réduire le nombre de motifs et donc d'accélérer l'extraction. Cela montre également que le temps de traitement de la contrainte est négligeable par rapport au temps gagné en élaguant l'espace de recherche. La figure 1.11 illustre cela sur les données de *Kwe Binyi* avec le modèle d'Atherton [Ath05] [2]. Ces résultats montrent notamment que des contraintes relativement "larges" ($minf = 3$ et surface minimum à 10% de la zone d'étude) permettent tout de même d'élaguer un grand nombre de motifs non intéressants.

Impact de l'approche de visualisation

Comme le montre la figure 1.12, l'extraction des co-localisations est plus lente lorsque celle-ci intègre le calcul des représentations visuelles, ce qui est normal car un *clustering* est fait en plus. Cependant, ces performances restent du même ordre de grandeur. Le nombre de motifs retournés aux experts est par contre totalement différent. En moyenne, chaque co-localisation est représentée par trois motifs sur la carte, alors que le nombre d'instances est beaucoup plus important.

1. Expérimentations réalisées sur un Intel (R) Xeon (R) 2.66GHz avec 4 Go de RAM et Windows Server 2003
2. Expérimentations réalisées sur un Intel (R) Xeon (R) 2.2GHz avec 8 Go de RAM et Windows Server 2012

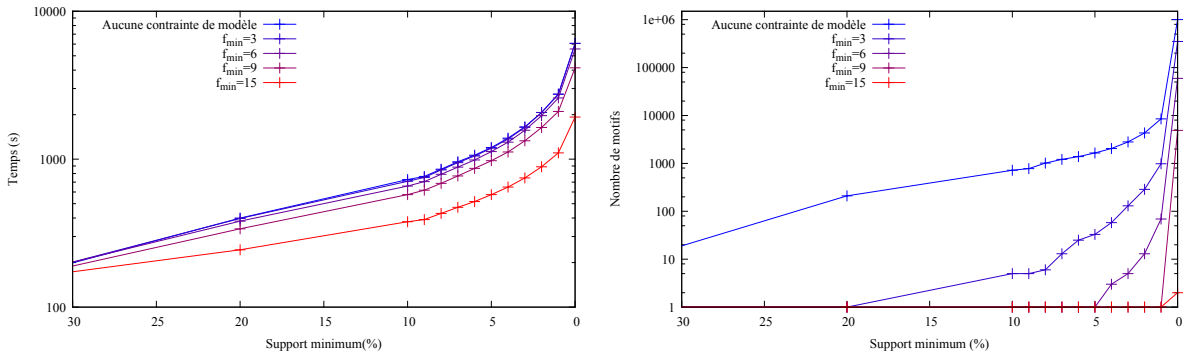


FIGURE 1.11 – Temps d'exécution et nombre de motifs extraits avec et sans contrainte basée sur le modèle d'Atherton (*Kwe Binyi*)

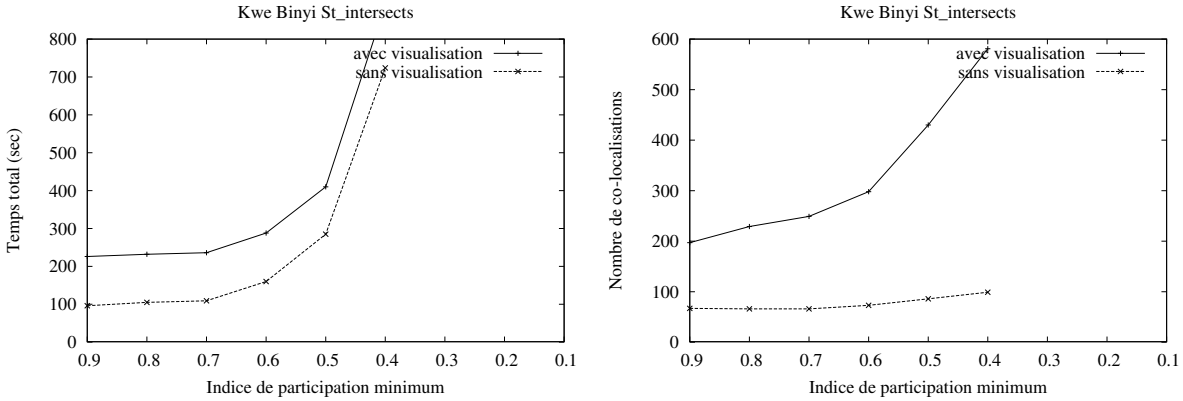


FIGURE 1.12 – Exemple de temps d'exécution et de nombre de motifs retournés avec et sans visualisation

En général, notre approche de *clustering* est plus rapide qu'un post-traitement basé sur *DBSCAN* [EKSX96] ou *X-means* [PM00]. Ce dernier peut toutefois être légèrement plus efficace pour des seuils élevés. Cette différence est principalement due au coût du pré-*clustering* effectué pour chaque type d'objets. En post-traitement, ces clustering ne sont pas effectués car les motifs de taille un ne sont pas retournés. Au contraire, nous devons les calculer et les stocker avec notre approche.

2

Extraction de motifs séquentiels intégrant le voisinage

L'extraction de co-localisations est une problématique importante en fouille de données spatiales. Comme discuté en section 2.2.1, bien que ce concept ait été étendu à des données spatio-temporelles [CSRS08, QHH09, PAA13, Cel15], l'intégration de la dimension temporelle dans les co-localisations reste structurellement limitée. D'autres approches ont été proposées telles que [TG01, YPM05, HZZ08, MSSR12, BTD17]. Toutefois, elles ne permettent pas de prendre en compte l'évolution d'un ensemble de types d'évènements tout en considérant l'environnement proche.

Ce constat a donc motivé une deuxième contribution visant à définir un nouveau domaine de motifs séquentiels intégrant le voisinage, et à développer un algorithme efficace pour les extraire. Ce travail a principalement été réalisé dans le cadre d'un projet de recherche du ministère des Outre-mer ("projet MOM") centré sur la dengue (une maladie vectorielle ayant un impact important en Nouvelle-Calédonie). Dans ce contexte, notre objectif était de développer des méthodes originales pour analyser la dynamique de la maladie dans les quartiers de Nouméa. A noter que ce travail a également été utilisé dans le cadre du projet ANR Fresqueau (ANRII-MONU14), mené par l'ENGEES à Strasbourg et TETIS à la Maison de la Télédétection de Montpellier pour l'étude de la qualité de l'eau des rivières.

Le tableau 2.1 présente les différents étudiants (stagiaire et doctorant) ayant travaillé sur cette problématique. Il présente aussi le projet de recherche et les collaborations associés à ce travail. Suite à ce tableau sont également listées les principales publications.

Master/Thèse
L. Mabit (2010) stage Université Lyon 2 co-encadrement N. Selmaoui-Folcher (UNC)
H. Alatrista Salas (2009-2012) thèse dirigée par M. Teisseire (IRSTEA) et N. Selmaoui-Folcher (UNC), co-encadrants : S. Bringay (Université de Montpellier) et F. Flouvat

Projet
projet MOM "Prévention et prédiction des épidémies de dengue en Nouvelle-Calédonie" (2010-2011)
Collaborations
IRSTEA, Université de Montpellier, IRD, Direction des Affaires Sanitaires et Sociales NC, Institut Pasteur NC, Météo NC

TABLE 2.1 – Synthèse des encadrements, des projets et des collaborations en lien avec l'extraction de motifs spatio-séquentiels

Ce chapitre est organisé de la manière suivante. La section 2.1 présente le domaine de motifs proposé et les contraintes utilisées. La section 2.2 décrit un algorithme dérivé de $PrefixSpan$

Principales publications
Nazha Selmaoui-Folcher and Frédéric Flouvat. How to use "classical" tree mining algorithms to find complex spatio-temporal patterns ? In Proceedings of the International Conference on Database and Expert Systems Applications (DEXA'11), Toulouse, France, August 2011.
Hugo Alatrista-Salas, Sandra Bringay, Frédéric Flouvat, Nazha Selmaoui-Folcher and Maguelonne Teisseire. The Pattern Next Door : Towards Spatio-Sequential Pattern Discovery. In Proceedings of the 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PaKDD'12), Kuala Lumpur, Malaysia, 2012.
Hugo Alatrista-Salas, Jérome Azé, Sandra Bringay, Flavie Cernesson, Frédéric Flouvat, Nazha Selmaoui-Folcher and Maguelonne Teisseire. Finding relevant sequences with the least temporal contradiction measure : application to hydrological data. In proceedings of the AGILE 2012 International Conference on Geographic Information Science, Avignon, April 24-27, pp197-202, 2012.
Hugo Alatrista Salas, Frédéric Flouvat, Sandra Bringay, Nazha Selmaoui-Folcher and Maguelonne Teisseire. A spatial-based KDD process to better manage the river water quality. In the International Journal of Geomatics and Spatial Analysis 23(3-4), p.469-494, 2013.
Hugo Alatrista Salas, Sandra Bringay, Frédéric Flouvat, Nazha Selmaoui-Folcher, Maguelonne Teisseire. Spatio-sequential patterns mining : beyond the boundaries. In Intelligent Data Analysis, Vol. 20(2), p.293-316, 2016.

TABLE 2.2 – Synthèse des publications en lien avec l'extraction de motifs spatio-séquentiels

pour extraire ces motifs. Une mesure de qualité visant à améliorer la pertinence des motifs affichés à l'expert est présentée en section 2.3. La section 2.4 présente un outil pour visualiser les motifs extraits. Pour finir, la section 2.5 discute des résultats obtenus sur un jeu de données lié à la dengue dans Nouméa.

2.1 Cadre théorique

2.1.1 Les données

Les données considérées dans ce chapitre décrivent la distribution d'attributs variant de manière continue dans l'espace et le temps (p.ex. conditions météorologiques, nombre de cas de dengue), mais qui sont échantillonnés et discrétisés pour être traités à l'échelle du quartier. Dans notre contexte, le modèle spatial considéré est le pavage de polygones (chacun représentant un quartier).

Une telle base de données peut être définie comme un triplet $BD = (d_S, d_T, d_A)$ où d_T est un ensemble d'estampilles temporelles, d_S est un ensemble de zones et d_A un ensemble de caractéristiques des zones au cours du temps (des valeurs catégorielles). Les zones sont liées par une relation de voisinage notée *voisin* définie par : $voisin(z_i, z_j) = vrai$ si z_i et z_j sont voisines, $faux$ sinon. Le tableau 2.3 présente une telle base de données liée aux données météorologiques de trois régions sur trois jours consécutifs. Les trois régions sont liées par une relation de proximité décrite dans la figure 2.1.

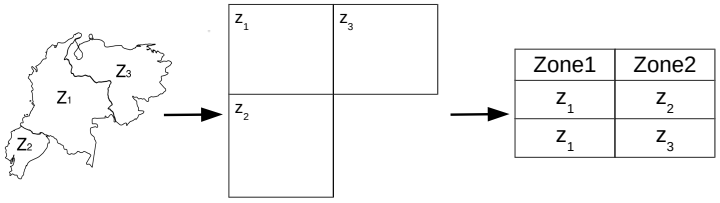


FIGURE 2.1 – Régions voisines

Région	Date	Température	Précipitation	Vent	Rafales
z_1	22/12/10	$temp_m$	$precip_m$	$vent_m$	[55,65[
z_1	23/12/10	$temp_m$	$precip_m$	$vent_l$	[75,100[
z_1	24/12/10	$temp_l$	$precip_m$	$vent_m$	[75,100[
z_2	22/12/10	$temp_m$	$precip_m$	$vent_m$	[55,65[
z_2	23/12/10	$temp_l$	$precip_m$	$vent_l$	[50,55[
z_2	24/12/10	$temp_l$	$precip_l$	$vent_m$	[75,100[
z_3	22/12/10	$temp_s$	$precip_s$	$vent_s$	[55,65[
z_3	23/12/10	$temp_m$	$precip_s$	$vent_l$	[55,65[
z_3	24/12/10	$temp_m$	$precip_s$	$vent_s$	[50,55[
...	...	...	...	...	...

TABLE 2.3 – Changements météorologiques dans trois régions : z_1, z_2 et z_3 pour le 22, 23, 24 décembre 2010

2.1.2 Les motifs spatio-séquentiels

Face à de telles données intégrant à la fois une dimension temporelle et spatiale, nous introduisons un nouveau domaine de motifs dérivé des motifs séquentiels et des co-localisations : les motifs spatio-séquentiels.

Ces motifs s'appuient sur le concept d'*itemset spatial*. Dans le cadre de ces données, un *itemset* IS est sous-ensemble de la dimension d'analyse d_A, i.e. $IS \subseteq d_A$. Un ***itemset* spatial** $I_{ST} = IS_i \odot IS_j$ représente deux *itemsets* proches spatialement. Autrement dit, il existe au moins une zone $z \in d_S$ associée à l'itemset IS_i à un temps $t \in d_T$, et cette zone est voisine d'une zone associée à l'itemset IS_j au même temps. Par exemple, $I_{ST} = \{temp_m, precip_m, vent_l\} \odot \{vent_l, [50, 55[\}$ est un *itemset spatial* associé à la zone z_1 le 23/12/10 car $\{vent_l, [50, 55[\}$ est au même moment associé à la zone voisine z_2.

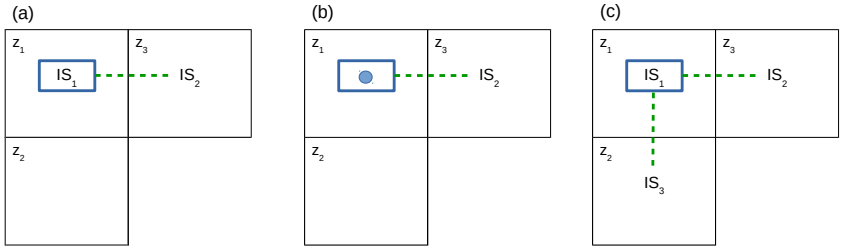


FIGURE 2.2 – Représentation graphique des itemsets spatiaux (a) $IS_1 \odot IS_2$ (b) $\theta \odot IS_2$ (c) $IS_1 \odot [IS_2; IS_3]$.

Afin de faciliter les notations, nous introduisons un **opérateur de groupement n-aire** noté [ ], qui permet de regrouper une liste d'*itemsets* affectés par l'opérateur $\odot$ (*à côté de*). Nous utilisons également le symbole θ pour représenter l'**absence d'itemsets** dans une zone. La figure 2.2 montre les trois types d'*itemsets* spatiaux que nous pouvons construire avec les notations introduites précédemment. Les lignes pointillées représentent la relation de voisinage spatiale. Par exemple, l'*itemset* spatial $I_{ST} = \theta \odot [\{temp_m\}; \{precip_l\}]$ représente une zone sans événement particulier mais dont deux zones voisines distinctes ont été marquées respectivement par une température $temp_m$ et des précipitations $precip_l$ au même moment.

Un itemset spatial $I_{ST} = IS_i \odot IS_j$ est **inclus** dans un autre itemset spatial $I'_{ST} = IS'_k \odot IS'_l$, noté $I_{ST} \subseteq I'_{ST}$, si et seulement si $IS_i \subseteq IS'_k$ et $IS_j \subseteq IS'_l$. Par exemple, l'itemset spatial

$I_{ST} = \{temp_m, precip_m\} \odot \{vent_l\}$ est *inclus* dans l'itemset spatial $I'_{ST} = \{temp_m, precip_m\} \odot \{vent_l, [55, 65[\}$ car $\{temp_m, precip_m\} \subseteq \{temp_m, precip_m\}$ et $\{vent_l\} \subseteq \{vent_l, [55, 65[\}$.

Dans ce contexte, un **motif spatio-séquentiel** est une liste ordonnée d'itemsets spatiaux, notée $S =< I_{ST_1}, I_{ST_2}, \ldots, I_{ST_m} >$ où $I_{ST_i}, I_{ST_{i+1}}$ respectent la contrainte de séquentialité temporelle pour tout $i \in [1..m-1]$. L'exemple de motif spatio-séquentiel illustré en figure 2.3 représente le motif $S =< \{temp_m\},\ \theta \odot [\{precip_l\}; \{vent_s\}],\ \{vent_l\} \odot [\{precip_l\}; \{temp_l\}] >$. Les flèches représentent la dynamique temporelle et les lignes pointillées représentent la relation de voisinage spatial entre *itemsets*.

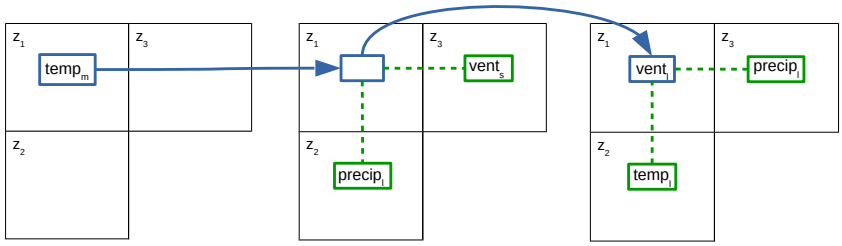


FIGURE 2.3 – Exemple de dynamique spatio-temporelle.

La relation d'inclusion définie pour les motifs séquentiels peut être étendue aux motifs spatio-séquentiels. Une séquence $S =< I_{ST_1}, I_{ST_2}, \ldots, I_{ST_m} >$ est **incluse** dans une séquence $S' =< I'_{ST_1}, I'_{ST_2}, \ldots, I'_{ST_n} >$, notée $S \preceq S'$, s'il existe des entiers $j_1 \leq \ldots \leq j_m$ tels que $I_{ST_1} \subseteq I'_{ST_{j_1}}, I_{ST_2} \subseteq I'_{ST_{j_2}}, \ldots, I_{ST_m} \subseteq I'_{ST_{j_m}}$. Par exemple, la séquence $S =< \{temp_l, precip_m\} \odot \{precip_l, vent_s\},\ \{[55, 65[\} >$ est incluse dans la séquence $S' =< \{temp_l, precip_m\} \odot \{precip_l, vent_s\},\ \{[55, 65[\} \odot \{vent_s\} >$ car $\{temp_l, precip_m\} \odot \{precip_l, vent_s\} \subseteq \{temp_l, precip_m\} \odot \{precip_l, vent_s\}$ et $\{[55, 65[\} \subseteq \{[55, 65[\} \odot \{vent_s\}$.

2.1.3 Les contraintes de fréquence et de participation minimales

Nous introduisons dans cette sous-section deux contraintes visant à filtrer des motifs spatio-séquentiels intéressants. Ces contraintes sont dérivées de la mesure de fréquence définie pour la fouille de motifs séquentiels [AS95] et de l'indice de participation défini pour la fouille de co-localisations spatio-temporelles [HSX04]. Afin d'illustrer ces deux mesures, nous utiliserons la base de séquences suivante issue de la transformation du tableau 2.4. Elle représente l'évolution des attributs pour chaque zone.

Région	**Séquence**
z_1	$S_{z_1} =< \{temp_m, precip_m, vent_m, [55, 65[\}, \{temp_m, precip_m, vent_l, [75, 100[\}, \{temp_l, precip_m, vent_m, [75, 100[\} >$
z_2	$S_{z_2} =< \{temp_m, precip_m, vent_m, [55, 65[\}, \{temp_l, precip_m, vent_l, [50, 55[\}, \{temp_l, precip_l, vent_m, [75, 100[\} >$
z_3	$S_{z_3} =< \{temp_s, precip_s, vent_s, [55, 65[\}, \{temp_m, precip_s, vent_l, [55, 65[\}, \{temp_m, precip_s, vent_s, [50, 55[\} >$

TABLE 2.4 – Base de séquences BD_{seq} illustrant l'évolution des zones z_1, z_2 et z_3

En fouille de séquences, la **fréquence (relative)** d'une sous-séquence est définie comme le nombre de séquences en entrée qui contiennent la sous-séquence, sur le nombre total de séquences en entrée. Cette définition peut être aisément étendue à notre contexte en intégrant simplement le voisinage spatial lors de la vérification de l'inclusion. Par exemple, la fréquence du motif $S =< \{precip_m\},\ \{temp_l\} \odot \{[75, 100[\} >$ dans la base de données BD présentée dans le tableau 2.4

est $supp(S, BD) = 2/3$. Cette mesure indique que le motif apparaît dans deux zones sur trois (les zones z_2 et z_1 comme illustré en figure 2.4).

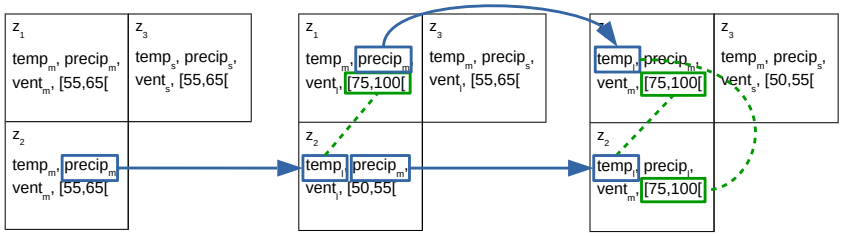


FIGURE 2.4 – Occurrences du motif $S =< \{precip_m\},\ \{temp_l\} \odot \{[75, 100[\} >$

On peut ainsi définir une contrainte de fréquence minimale similaire à celle classiquement utilisée dans la littérature. Soient $supp(S, BD)$ la fréquence relative de la séquence spatiale S dans la base de données BD et un seuil $minsup$ dans $[0, 1]$, la **contrainte de fréquence minimale** sur S est définie de la manière suivante :

$$Q_{supp}(S, BD) \equiv supp(S, BD) \geq minsup$$

L'indice de participation défini dans [HSX04] peut aussi être adapté en combinant un **indice de participation spatiale** et un **indice de participation temporelle**. L'indice de participation spatiale d'une séquence spatiale S, noté $SPi(S, BD)$, correspond à la probabilité minimale d'avoir une zone contenant S sachant un item $i \in S$. L'**indice de participation temporelle** de S, noté $TPi(S, BD)$, correspond à la probabilité minimale d'avoir une occurence d'un item de S sachant toutes les occurrences de ce même *item* dans une séquence en entrée.

$$SPi(S, BD) = min_{\forall f \in dom(d_A), f \in S} \left(\frac{supp(S, BD)}{supp(f, BD)} \right)$$

$$TPi(S, BD) = min_{S_z \in BD_{seq}, z \in dom(d_S)} \left(min_{f \in dom(d_A)} \left(\frac{\text{nb occurrences de } f \text{ dans } S}{\text{nb occurrences de } f \text{ dans } S_z} \right) \right)$$

Par exemple, le motif $S =< \{precip_m\},\ \{temp_l\} \odot \{[75, 100[\} >$ de la figure 2.4 a un indice de participation spatiale de $2/2 = 1$ et un indice de participation temporelle de $1/3$. L'**indice de participation spatio-temporelle** d'une séquence spatiale s, noté $STPi(S, BD)$, est ensuite défini comme le produit des deux mesures précédentes, i.e. $STPi(S, BD) = SPi(S, BD) * TPi(S, BD)$. Dans l'exemple précédent, le motif S a donc un indice de participation spatio-temporelle de $1/3$.

Comme précédemment, on peut définir une **contrainte de participation minimale** de la manière suivante :

$$Q_{stpi}(S, BD) \equiv STPi(S, BD) \geq minstpi$$

Notons que ces deux prédicats $Q_{supp}(S, BD)$ et $Q_{stpi}(S, BD)$ sont anti-monones. Si un motif spatio-séquentiel S n'est pas "fréquent" ou intéressant, alors tous les motifs S' tel que $S \preceq S'$, le sont également. Ces contraintes peuvent donc être utilisées dans les algorithmes d'extraction pour élaguer l'espace de recherche.

2.1.4 Problématique

Soit une base de données spatio-temporelles BD. L'objectif est de trouver l'ensemble des motifs spatio-séquentiels dont la fréquence (ou l'indice de participation spatio-temporelle) est supérieure à un seuil spécifié par l'utilisateur.

2.2 Stratégie d'extraction des motifs spatio-séquentiels

Etant donné la structure de ce domaine de motifs et les propriétés de ces contraintes, l'extraction peut être effectuée suivant une stratégie dérivée de celles utilisées pour extraire les séquences fréquentes. Les principales modifications résident dans la façon de générer les motifs et dans les calculs effectués pour vérifier la contrainte choisie. Deux stratégies ont été plus particulièrement adaptées. La première est la stratégie utilisée par l'algorithme *Apriori* [AS95] (parcours par niveau) et la deuxième est la stratégie utilisée par l'algorithme $PrefixSpan$ [PHMA$^+$01] (parcours en profondeur et projections successives).

2.2.1 Algorithme par niveau dérivé d'*Apriori*

Pour rappel, le principe de cette stratégie d'exploration est d'effectuer un parcours en largeur (encore appelé par niveau) de l'espace de recherche. Tout d'abord, l'algorithme recherche des motifs intéressants de taille un. Ensuite, à chaque itération k, un ensemble de motifs candidats de taille k, noté $Cand_k$, est généré en utilisant des motifs intéressants de taille $k-1$. Un motif candidat est un motif dont tous les sous-modèles de taille $k-1$ sont intéressants. Après ces étapes de génération et d'élagage, tous les motifs candidats sont testés par rapport au prédicat, et les motifs intéressants sont utilisés pour commencer la prochaine itération. L'algorithme s'arrête lorsque l'ensemble des motifs candidats est vide.

Cette stratégie initialement développée pour des *itemsets* a été étendue aux séquences (d'*itemsets*) dans [AS95] en rajoutant de nouvelles règles de génération des motifs candidats. L'intégration des *itemsets* spatiaux dans les motifs séquentiels nécessite donc d'étendre ces règles afin de pouvoir générer tous les motifs de ce langage plus riche (cinq règles ont dû être ajoutées). L'ensemble des règles de génération sont décrites dans le tableau 2.5. Le motif $< \rho x >$ représente une séquence dont le préfixe est la séquence ρ et dont le dernier *itemset* finit par l'item x. Le motif $< \rho, \{x\} >$ représente quant à lui une séquence dont le préfixe est la séquence ρ suivi par un *itemset* $\{x\}$.

Table 2.5 – Règles de génération des motifs candidats

Séquence α	**Séquence β**	**Si**	**Motif candidat γ**
$< \rho x >$	$< \rho y >$	$x < y$	$< \rho xy >$
$< \rho x >$	$< \rho, \{y\} >$		$< \rho x, \{y\} >$
$< \rho, \{x\} >$	$< \rho, \{y\} >$	$x < y$	$< \rho, \{x, y\} >$
$< \rho x >$	$< \rho \odot y >$		$< \rho x \odot y >$
$< \rho, \{x\} >$	$< \rho \odot y >$		$< \rho \odot y, \{x\} >$
$< \rho \odot x >$	$< \rho \odot y >$	$x < y$	$< \rho \odot [x; y] >$ et $< \rho \odot \{x, y\} >$
$< \rho x >$	$< \rho, \theta \odot \{y\} >$		$< \rho x, \theta \odot \{y\} >$
$< \rho, \theta \odot \{x\} >$	$< \rho, \theta \odot \{y\} >$	$x < y$	$< \rho, \theta \odot [x; y] >$ et $< \rho, \theta \odot \{x, y\} >$

Pour calculer l'intérêt d'un motif spatio-séquentiel, il est nécessaire d'étudier l'évolution de chaque zone mais aussi celles des zones voisines. Afin de faciliter les tests, la base de données est

stockée sous forme séquentielle dans une structure de données simple divisée en deux parties : la première stocke les séquences représentant l'évolution de chaque zone étudiée, et la deuxième référence les séquences associées à l'évolution des zones voisines (cf. tableau 2.6).

Séquence	**Séquences voisines**
$S_{z_1} =< \{temp_m, precip_m, vent_m, [55,65[\}, \{temp_m, precip_m, vent_l, [75,100[\}, \{temp_l, precip_m, vent_m, [75,100[\} >$	S_{z_2}, S_{z_3}
$S_{z_2} =< \{temp_m, precip_m, vent_m, [55,65[\}, \{temp_l, precip_m, vent_l, [50,55[\}, \{temp_l, precip_l, vent_m, [75,100[\} >$	S_{z_1}
$S_{z_3} =< \{temp_s, precip_s, vent_s, [55,65[\}, \{temp_m, precip_s, vent_l, [55,65[\}, \{temp_m, precip_s, vent_s, [50,55[\} >$	S_{z_1}
...	...

TABLE 2.6 – Stockage des données en entrée intégrant le voisinage

2.2.2 Algorithme en profondeur dérivé de *PrefixSpan*

Comme l'ont montré les expérimentations, le passage à l'échelle de l'approche par niveau reste limité lorsque les motifs solutions sont longs. Face à ce problème, nous avons testé une stratégie basée sur $PrefixSpan$ [PHMA+01]. Pour rappel, le principe de cette approche est d'extraire les solutions sans génération de motifs candidats. Cette approche fait récursivement des projections de la base de données, recherche des extensions fréquentes et re-projette la base de données en fonction de celles-ci. Les motifs fréquents sont ainsi étendus progressivement suivant un parcours en profondeur de l'espace de recherche.

Cette approche est décrite par l'algorithme 2. Cette procédure commence par rechercher l'ensemble Th_1 des *items* f et $\theta \odot \{f\}$ vérifiant Q dans la projection de la base de données BD par rapport au préfixe α, noté $BD|_\alpha$ (ligne 1 de l'algorithme 2). Ces *items* vont constituer les extensions possibles de la séquence α. Ensuite, pour chaque *item* ou *item* spatial $x \in Th_1$, nous étendons la séquence spatiale α avec x (lignes 3-4). Pour faire cela, nous avons deux possibilités : 1) ajouter x au dernier *itemset* spatial de la séquence α (ligne 3) ou 2) insérer x après (i.e. au temps suivant) le dernier *itemset* spatial de α (ligne 4). Une fois l'extension du motif effectuée, nous vérifions si les séquences spatiales résultantes vérifient la contrainte Q (lignes 5 et 9). Chaque solution trouvée est enregistrée dans l'ensemble $Th(BD, Q)$ (lignes 6 et 10). L'algorithme effectue ensuite une projection de la base de données par rapport à ces motifs. Les appels récursifs qui suivent permettent de construire les sur-séquences des motifs fréquents trouvés à partir des bases projetées correspondantes (lignes 7 et 11). L'algorithme s'arrête lorsqu'il n'y a plus de projections qui puissent être générées.

Nous illustrons cet algorithme en utilisant des données du tableau 2.3 et la figure 2.1 avec une fréquence minimale de 2/3. Dans un premier temps, *Prefix-growthST* commence par extraire les *items* fréquents et *items* spatiaux fréquents de BD (ligne 1), soit : $Th_1 = \{precip_m, temp_l, temp_m, vent_m, vent_l, [50,55[, [55,65[, [75,100[, \theta \odot precip_m, \theta \odot temp_l, \theta \odot temp_m, \theta \odot vent_l, \theta \odot vent_m, \theta \odot [55,65[, \theta \odot [75,100[\}$.

Pour chaque *item* fréquent f et $\theta \odot f$ trouvé (ligne 2), l'algorithme calcule la projection de la base de données par rapport à cet *item* (aucune extension n'est faite ici car α est vide). Par exemple, pour l'item fréquent $precip_m$, nous obtenons la projection présentée dans le tableau 2.7.

Chacune de ces bases projetées est ensuite utilisée dans un appel récursif permettant de rechercher des sur-séquences solutions (lignes 7 et 11). Le premier appel récursif va construire les sur-séquences ayant pour préfixe $< \{precip_m\} >$ à partir de la base projetée décrite dans le tableau 2.7. Plus précisément, l'algorithme va d'abord rechercher les *items* fréquents dans cette base projetée (lignes 1), puis étendre $< \{precip_m\} >$ avec. Les items fréquents obtenus

Algorithm 2 Prefix-growthST($\alpha,\ BD|_\alpha,\ Q,\ Th(BD,Q)$)

Require: une séquence spatiale α , la projection $BD|_\alpha$ de la base de données spatio-temporelles par rapport à α, une contrainte anti-monotone Q, et $Th(BD,Q)$ un ensemble de motifs spatio-séquentiels solutions ;

1: $Th_1 \leftarrow$ {l'ensemble d'*items* f et $\theta \odot \{f\}$ vérifiant Q dans $BD|_\alpha$, avec $f \in dom(d_A)$ }
2: **for all** $x \in Th_1$ **do**
3: $\quad \beta \leftarrow < \alpha x >$
4: $\quad \delta \leftarrow < \alpha, \{x\} >$
5: $\quad$ **if** $Q(\beta, BD|_\alpha)$ **then**
6: $\quad\quad Th(BD,Q) \leftarrow Th(BD,Q) \cup \beta$;
7: $\quad\quad$ Prefix-growthST$(\beta, BD|_\beta, Q, Th(BD,Q))$
8: $\quad$ **end if**
9: $\quad$ **if** $Q(\delta, BD|_\alpha)$ **then**
10: $\quad\quad Th(BD,Q) \leftarrow Th(BD,Q) \cup \delta$;
11: $\quad\quad$ Prefix-growthST($\delta,\ BD|_\delta,\ Q,\ Th(BD,Q)$)
12: $\quad$ **end if**
13: **end for**

Séquence	**Séquences voisines**
$S_{z_1} =< \{\_, vent_m, [55,65[\}, \{temp_m, precip_m, vent_l, [75,100[\}, \{temp_l, precip_m, vent_m, [75,100[\} >$	$S_{z_2} =< \{\_, vent_m, [55,65[\}, \{temp_l, precip_m, vent_l, [50,55[\}, \{temp_l, precip_l, vent_m, [75,100[\} >$ $S_{z_3} =< \{\_, vent_s, [55,65[\}, \{temp_m, precip_s, vent_l, [55,65[\}, \{temp_m, precip_s, vent_s, [50,55[\} >$
$S_{z_2} =< \{\_, precip_m, vent_m, [55,65[\}, \{temp_l, precip_m, vent_l, [50,55[\}, \{temp_l, precip_l, vent_m, [75,100[\} >$	$S_{z_1} =< \{\_, precip_m, vent_m, [55,65[\}, \{temp_m, precip_m, vent_l, [75,100[\}, \{temp_l, precip_m, vent_m, [75,100[\} >$

TABLE 2.7 – Projection de $< \{precip_m\} >$ sur BD

pour $BD|_{<(precip_m\}>}$ sont $\{vent_l, vent_m, temp_l, precip_m, [55,65[, [75,100[, \theta \odot vent_l, \theta \odot vent_m, \theta \odot temp_l, \theta \odot precip_m, \theta \odot [55,65[, \theta \odot [75,100[\}$. Le premier *item* fréquent trouvé est $vent_l$. On obtient donc les motifs $< \{precip_m, vent_l\} >$ (ligne 3) et $< \{precip_m\}, \{vent_l\} >$ (ligne 4). Seul $< (precip_m\}, \{vent_l\} >$ est fréquent (ligne 9), l'algorithme utilise ce motif pour faire une nouvelle projection (cf. tableau 2.8) et rechercher récursivement toutes les sous-séquences fréquentes ayant pour préfixe $< \{precip_m\}, \{vent_l\} >$.

Séquence	**Séquences voisines**
$S_{z_1} =< \{\_, [75,100[\}, \{temp_l, precip_m, vent_m, [75,100[\} >$	$S_{z_2} =< \{\_, [50,55[\}, \{temp_l, precip_l, vent_m, [75,100[\} >$ $S_{z_3} =< \{\_, [55,65[\}, \{temp_m, precip_s, vent_s, [50,55[\} >$
$S_{z_2} =< \{\_, [50,55[\}, \{temp_l, precip_l, vent_m, [75,100[\} >$	$S_{z_1} =< \{\_, [75,100[\}, \{temp_l, precip_m, vent_m, [75,100[\} >$

TABLE 2.8 – Projection de $< \{precip_m\}, \{vent_l\} >$ sur BD

L'*item* spatial $\theta \odot temp_l$ est un des *items* fréquents dans l'appel récursif qui suit. L'algorithme construira donc ensuite les motifs $< \{precip_m\}, \{vent_l\} \odot \{temp_l\} >$ et $< \{precip_m\}, \{vent_l\}, \theta \odot \{temp_l\} >$. Une fois tous les *items* fréquents projetés, l'algorithme va parcourir une autre "branche" de l'espace de recherche (les motifs commençant par $< \{temp_l\} >$ par exemple). L'algorithme procède donc globalement de la même manière, que les *items* soient spatiaux ou non. Notons que lorsque l'algorithme étend un motif du type $< I_{ST_1}, I_{ST_2}, ..., IS_{ST_k} \odot \{x\} >$ avec un *item* fréquent $\theta \odot y$ $(x < y)$, la ligne 3 construit les motifs $< I_{ST_1}, I_{ST_2}, ..., IS_{ST_k} \odot \{x, y\} >$ et $< I_{ST_1}, I_{ST_2}, ..., IS_{ST_k} \odot \{x\} \odot \{y\} >$. Dans ce dernier cas, l'opérateur de groupement n-aire est utilisé afin de représenter la séquence sous la forme $< I_{ST_1}, I_{ST_2}, ..., IS_{ST_k} \odot [x; y]\} >$.

2.3 Evaluation de la qualité des motifs extraits

A l'issue de l'extraction, un grand nombre de motifs spatio-séquentiels peuvent être extraits. Afin d'éliminer les motifs non pertinents, nous avons introduit une nouvelle mesure de qualité basée sur le concept de moindre contradiction proposé dans [Azé03]. Cette mesure calculée en post-traitement vise à filtrer les motifs les plus contradictoires par rapport aux données.

Le concept de moindre contradiction a été introduit dans le cadre des règles d'association. Soit une règle d'association $X \rightarrow Y$, avec X et Y deux *itemsets* disjoints. La mesure de moindre contradiction de cette règle est $MC(X \rightarrow Y) = \frac{supp(XY)-supp(X\overline{Y})}{supp(Y)}$, où $supp(X\overline{Y})$ représente les transactions contenant X mais pas Y. Cette mesure est relativement proche de la mesure de *lift* communément utilisée pour filtrer les règles d'association : $lift(X \rightarrow Y) = \frac{supp(XY)}{supp(X)\times supp(Y)}$. Elle est appliquée en post-traitement sur toutes les règles extraites. L'avantage de la mesure de moindre contradiction est d'être facile à interpréter. De plus, comme montré dans [Azé03], elle permet d'extraire des connaissances intéressantes et elle résiste relativement bien au bruit.

Une première étape avant de pouvoir l'appliquer aux motifs spatio-séquentiels consiste à l'adapter aux séquences d'*itemsets*. Soit une sous-séquence S apparaissant dans une base de données séquentielle BD_{seq}. Nous considérons que cette séquence S est "contredite" par une autre séquence S_c, si cette dernière contient les mêmes *itemsets* que S mais dans un ordre différent. Considérons les ensembles $\mathcal{S}_{Contr}$ et $\mathcal{S}_{All}$, représentant respectivement l'ensemble des sous-séquences extraites contenant tous les *itemsets* de S dans un ordre différent et l'ensemble des sous-séquences extraites contenant tous les *items* de S. La **mesure de moindre contradiction temporelle** de S dans BD_{seq} est

$$MCT(S, BD_{seq}) = \frac{supp(S, BD_{seq}) - \sum\limits_{\min_{\subseteq}\{S_c \in \mathcal{S}_{Contr}\}} supp(S_c, BD_{seq})}{\sum\limits_{\min_{\subseteq}\{S_a \in \mathcal{S}_{All}\}} supp(S_a, BD_{seq})}$$

L'exemple suivant présente une partie des solutions extraites dans une base de données séquentielles BD_{seq} (avec leur fréquence) et la mesure de moindre contradiction temporelle d'un de ces motifs.

$Th(BD, Q) = \{$
$S_0 =< \{b, c\}, \{a\}, \{d\} >$, $supp(S_0, BD_{seq}) = 0.14$
$S_1 =< \{a\}, \{b, c\} >$, $supp(S_1, BD_{seq}) = 0.35$
$S_2 =< \{b, c\}, \{a\} >$, $supp(S_2, BD_{seq}) = 0.15$
$S_3 =< \{a, b\}, \{c, e\} >$, $supp(S_3, BD_{seq}) = 0.30$
$S_4 =< \{a, e\}, \{b, d\} >$, $supp(S_4, BD_{seq}) = 0.20$
$S_5 =< \{a, b\}, \{b\} >$, $supp(S_5, BD_{seq}) = 0.26$
...
$\}$

$$MCT(S_1, BD) = \frac{0.35 - (0.15)}{(0.35 + 0.15 + 0.30)} = 0.25$$

avec $\mathcal{S}_{Contr} = \{S_2\}$
et $\mathcal{S}_{All} = \{S_1, S_2, S_3\}$

FIGURE 2.5 – Exemple de moindre contradiction temporelle

Comme le montre cet exemple, nous ne prenons pas en compte tous les motifs "contradictoires" car certains d'entre eux sont redondants. Notamment, nous ne considérons pas la contradiction exprimée par le motif S_0 car celle-ci est redondante par rapport à celle exprimée par S_2. En effet, S_2 est inclus dans S_0. Ce dernier apparaît donc dans une partie des séquences en entrée contenant S_2. Il représente donc aussi une partie des contradictions de S_2. Ceci explique pourquoi nous ne conservons que les plus petits motifs contradictoires par rapport à l'inclusion

($\min_{\subseteq}$ dans l'équation).

Dans l'absolu, il est possible d'appliquer directement cette définition aux séquences spatiales, car la mesure de fréquence définie en sous-section 2.1.3 intègre déjà le voisinage. Toutefois, cette définition considérerait différemment les *itemsets* spatiaux $IS_i \odot IS_j$ et $IS_j \odot IS_i$, alors qu'ils représentent la même chose (la relation de voisinage étant symétrique). Nous adaptons donc la définition précédente afin de prendre en compte ces symétries. La **mesure de moindre contradiction spatio-temporelle** de la séquence spatiale S dans BD est

$$MCST(S, BD) = \frac{supp(S, BD) + \sum_{\min_{\subseteq}\{S_s \in \mathcal{S}_{Sim}\}} supp(S_s, BD) - \sum_{\min_{\subseteq}\{S_c \in \mathcal{S}_{Contr}\}} supp(S_c, BD)}{\sum_{\min_{\subseteq}\{S_a \in \mathcal{S}_{All}\}} supp(S_a, BD)}$$

avec
- $\mathcal{S}_{Sim}$ l'ensemble des séquences spatiales de BD incluant tous les *itemsets* spatiaux similaires de la séquence S dans le même ordre
- $\mathcal{S}_{Contr}$ l'ensemble des séquences spatiales de BD incluant tous les *itemsets* spatiaux similaires de la séquence S mais dans un ordre différent
- $\mathcal{S}_{All}$ l'ensemble des séquences de BD incluant tous les *items* qui sont apparus dans la séquence S

L'exemple suivant présente une partie des solutions extraites dans une base de données spatio-séquentielles BD (avec leur fréquence) et la mesure de moindre contradiction temporelle d'un de ces motifs.

$Th(BD, Q) = \{$
$S_0 =< \{a\}, \{b, c\} \odot \{f\} >$, $supp(S_1, BD) = 0.22$
$S_1 =< \{a\}, \{f\} \odot \{b, c\} >$, $supp(S_1, BD) = 0.35$
$S_2 =< \{b, c\} \odot \{f\}, \{a\} >$, $supp(S_2, BD) = 0.15$
$S_3 =< \{a, b\}, \{c\} \odot \{e\} >$, $supp(S_3, BD) = 0.20$
$S_4 =< \{a, f\}, \{b, c\} >$, $supp(S_4, BD) = 0.30$
$S_5 =< \{f\} \odot \{b, c\}, \{b\}, \{a, c\} >$, $supp(S_5, BD) = 0.26$
...
$\}$

$$MCST(S_1, BD) = \frac{0.35 + (0.22) - (0.15 + 0.26)}{(0.22 + 0.35 + 0.15 + 0.30 + 0.26)} = 0.125$$

avec $\mathcal{S}_{Contr} = \{S_2, S_5\}$
et $\mathcal{S}_{All} = \{S_0, S_1, S_2, S_4, S_5\}$

FIGURE 2.6 – Exemple de moindre contradiction spatio-temporelle

2.4 Visualisation des motifs spatio-séquentiels

L'approche choisie pour visualiser les séquences extraites est plus abstraite que celle proposée pour les co-localisations. Elle s'appuie sur une représentation sous forme de graphe. Elle n'intègre pas les principes et mécanismes des *SIG* car cet outil n'était pas au coeur des pratiques des experts côtoyés dans le cadre de notre application sur l'étude de la dengue. Pour autant, l'outil développé prend en compte la dimension cartographique des données. De manière générale, nous avons opté pour une visualisation plus classique basée sur une liste de motifs à partir de laquelle l'utilisateur peut sélectionner un motif qu'il souhaite étudier plus en détail.

2.4.1 Représentation visuelle d'un motif spatio-séquentiel

L'objectif de cette représentation sous forme de graphe est de mettre en avant la dynamique temporelle tout en prenant en compte les corrélations spatiales. La figure 2.7 illustre cette représentation sur l'exemple de la séquence spatiale $< \{a, b\}, \theta \odot [\{b\}; \{c\}], \{p\} \odot [\{q\}; \{r\}; \{d, e\}] >$. Dans cette figure, chaque noeud représente un *itemset* et chaque arête orientée représente la relation de précédence temporelle entre deux *itemsets*. Les arêtes en pointillé représentent les *itemsets* apparaissant dans une zone voisine. La taille des noeuds est proportionnelle au nombre d'*items* contenus dans l'*itemset*. De plus, l'utilisateur a la possibilité de mettre en avant certains évènements (ou valeurs) d'intérêt en leur associant une couleur (p.ex. "rouge" pour b dans la figure).

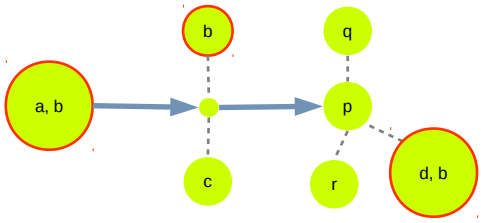


FIGURE 2.7 – Représentation graphique du motif $< \{a, b\}, \theta \odot [\{b\}; \{c\}], \{p\} \odot [\{q\}; \{r\}; \{d, e\}] >$

Un même graphe peut être affiché de différentes façons dans un plan. Notre approche de visualisation permet d'afficher un même motif selon quatre modèles : "fixe", "force", "arc", et "spiral" (cf. figure 2.8). La représentation précédente correspond à un affichage fixe, de gauche à droite, le long d'un axe horizontal (représentant le temps). Elle est probablement la plus facile à interpréter. Toutefois, si le motif est long, il ne sera pas possible de l'afficher intégralement à l'écran. A l'opposé, il est possible d'afficher un graphe selon un algorithme de force [KW03], i.e. de façon à ce que les arêtes soient environ de même longueur et qu'elles se croisent le moins possible. Cette représentation permet d'afficher un grand graphe dans un espace réduit et même de déplacer dynamiquement ses noeuds, mais elle reste plus difficile à interpréter. Les affichages en "arc" et spiral" permettent de représenter le motif selon un arc représentant le temps ou selon une spirale d'Archimède si le motif est trop long. Le choix du modèle d'affichage est laissé à l'utilisateur car il dépend beaucoup de ses préférences et du motif étudié.

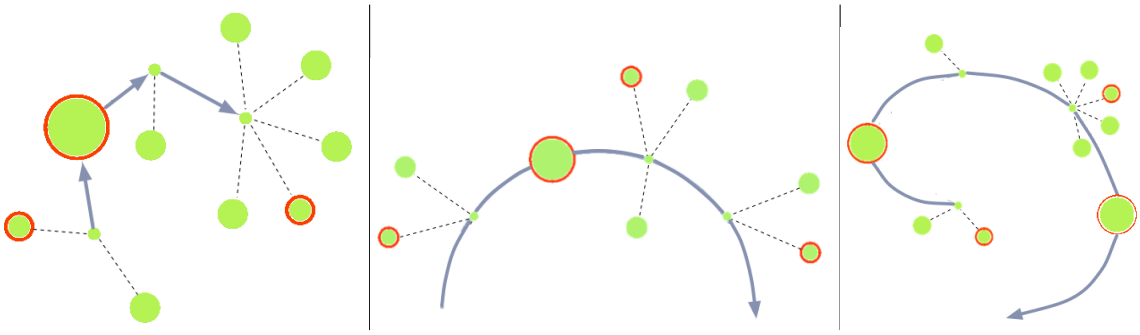


FIGURE 2.8 – Les autres modèles d'affichage d'une séquence spatiale (de gauche à droite : "force", "arc" et "spiral")

Au delà des informations portées directement par le motif, il est important pour les utilisateurs de savoir où et quand celui-ci est apparu. Pour cette raison, la représentation graphique d'un motif est suivie d'un tableau précisant les zones et les temps où il a eu lieu (cf. figure 2.9).

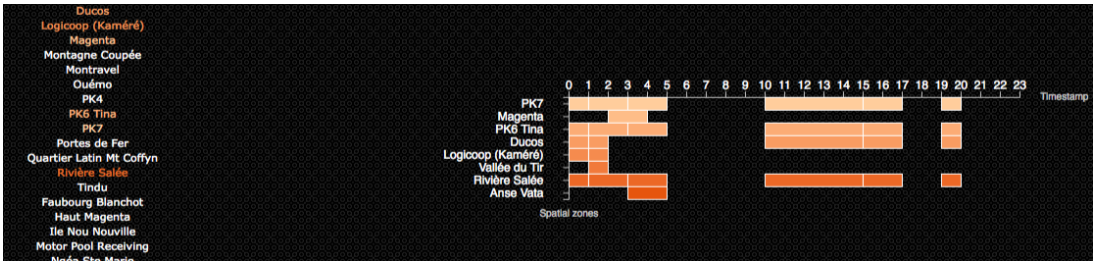


FIGURE 2.9 – Visualisation des zones et des temps associés à un motif

2.4.2 Visualisation de l'ensemble des solutions

Pour restituer l'ensemble des solutions extraites, un affichage classique basé sur une double visualisation est adopté. D'un côté, les zones géographiques associées aux données (les quartiers de Nouméa dans la figure 2.10) sont affichées sur un fond cartographique. De l'autre, une liste textuelle restitue les solutions avec leur mesure d'intérêt. Lorsque l'utilisateur sélectionne une ou plusieurs zones sur la carte, la liste des solutions est actualisée afin de ne présenter que les solutions associées aux zones sélectionnées. Inversement, lorsqu'un motif est sélectionné dans la liste, les zones et les temps associés à ce motifs sont affichés (avec la représentation graphique du motif). Par ailleurs, l'utilisateur peut filtrer les motifs par rapport à des évènements d'intérêt et étudier différents seuils de contrainte.

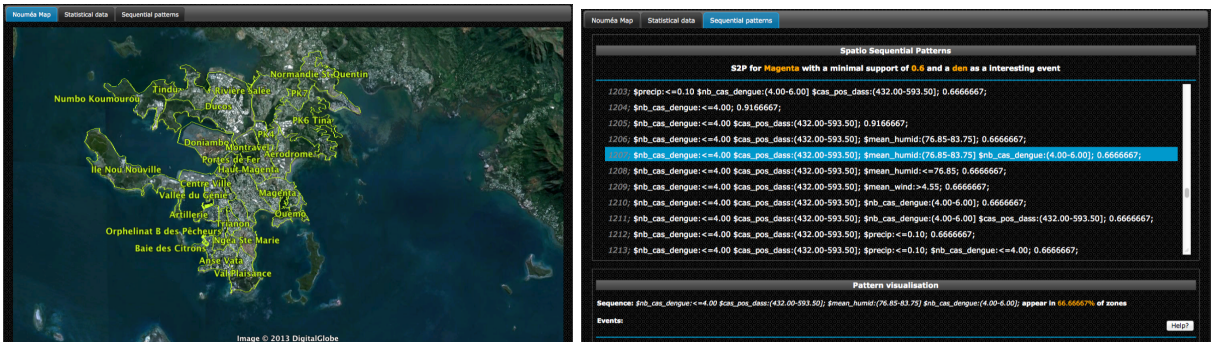


FIGURE 2.10 – Visualisation des zones et des motifs associés

2.5 Expérimentations et application au suivi de la dengue

2.5.1 Prototype

Le prototype développé se décompose en deux parties. La partie extraction contient les algorithmes présentés précédemment pour extraire les motifs spatio-séquentiels. Ces algorithmes ont été implémentés en Java. La partie visualisation correspond à une interface Web utilisée pour afficher et parcourir les motifs extraits. Cette interface est basée sur JQuery, la librairie D3 et Google Earth.

2.5.2 Protocole expérimental

Nous avons utilisé notre algorithme et la mesure de moindre contradiction pour étudier l'évolution de la dengue dans 12 quartiers de Nouméa en 2003 (24 dates sont considérées, corres-

pondant à des agrégations à la semaine). Pour chaque quartier, nous avons des informations démographiques (nombre d'habitants, de ménages, etc.), entomologique (p.ex. indice Breteau, indice de haut risque, etc.), météorologiques (précipitations, force du vent, température, etc.), d'urbanisation (nombre de points d'eau, d'écoles, de poubelles, etc.) et le nombre de cas de dengue. Au total, nous avons 11 attributs d'analyse. Pour chaque attribut, les données ont été discrétisées par fréquence égale en trois classes.

En plus de ce jeu de données, nous avons généré des jeux de données synthétiques afin d'évaluer plus en détail les performances de nos propositions. Ces jeux de données sont construits aléatoirement en fonction des paramètres suivants : nombre de zones, type de voisinage, nombre de dates et nombre moyen d'*items* par zone. Les zones et leur voisinage peuvent être définis selon deux configurations de l'espace : grille (huit voisins par zone) ou graphe (quatre voisins par zone). Au final, cinq jeux de données synthétiques ont été étudiés avec un nombre de zones variant de 10 à 20, un nombre de dates variant de 50 à 100, et un nombre moyen d'*items* par zone de 5.

Ces jeux de données ont été utilisés pour comparer les deux stratégies d'explorations proposées. Nous avons aussi étudier l'impact des paramètres suivants sur l'extraction : le type de voisinage, le nombre de zones, le nombre de dates, la densité et la contrainte utilisée (fréquence ou indice de participation). Les résultats ont été analysés d'un point de vue qualitatif et quantitatif. Les motifs extraits ont notamment été comparés à ceux extraits par l'approche de [TG01] n'intégrant pas le voisinage dans les motifs séquentiels.

2.5.3 Analyse qualitative

Le tableau 2.9 présente des exemples de motifs extraits liés à la dengue (seuil minimum de fréquence à 0.6). Par exemple, le dernier motif représente une évolution apparaissant dans plus de la moitié des zones (60%). Elle correspond à de faibles précipitations et quelques zones de stockage d'eau dans deux zones voisines, suivi par des cas de dengue dans la zone étudiée, suivi par le développement des gîtes larvaires des moustiques. Les zones vérifiant ces motifs sont des zones à risques car il y a une relation connue entre précipitations, gîtes larvaires, et propagation de la dengue. Toutefois, on constate que cette séquence est souvent contredite (indice de moindre contradiction égal à -0.04). A l'opposé, le premier motif apparaissant dans autant de zones, et représentant le lien entre gîtes larvaires et dengue, est très peu contredit (0.8). Cet exemple met en avant l'intérêt de la mesure de moindre contradiction spatio-temporelle afin de "pondérer" la pertinence des motifs extraits.

Motifs	**Fréquence**	**MCST**
< {ihre_index :>34.82, nb_cas_dengue :<=6.00} >	0.6	0.8
< {mean_wind :<=3.20}, {comunity_gather :<=20.00 }⊙[waste_container :<=39.00 ; nb_cas_dengue :<=6.00]>	0.76	0.54
< $\theta \odot$ { ihre_index :>34.82}, $\theta\odot$ [nb_cas_dengue :<=6.00 ; ihre_index :<=24.55], $\theta \odot$ {comunity_gather :<=20.00 }, {nb_cas_dengue :<=6.00, mean_temper :<=23.55} >	0.64	0.32
< $\theta\odot$[precip :<=0.10 ; indoor_deposit :2126-2692], {nb_cas_dengue :<=6.00 }, $\theta \odot$ {ihre_index :<=24.55} >	0.6	−0.04
...	...	...

TABLE 2.9 – Exemples de motifs spatio-séquentiels extraits dans les données "dengue"

Plusieurs motifs d'intérêts pour les experts ont été mis en avant grâce aux motifs extraits. On notera notamment le lien entre dengue et degré d'urbanisation, mis en avant par le biais de motifs intégrant le nombre de lampadaires dans les quartiers.

2.5.4 Analyse quantitative

Les jeux de données synthétiques ont ensuite été utilisés pour évaluer le passage à l'échelle des différentes approches en fonction des caractéristiques des données. Les expérimentations ont été réalisées sur un PC basé sur Intel (R) Xeon (R) avec 16 Go de RAM et Ubuntu Server 9.10 comme système d'exploitation.

Comme le montre la figure 2.11, l'algorithme en profondeur dérivé de *PrefixSpan* [PHMA+01] est plus performant que celui dérivé d'*Apriori* [AS95]. Ce comportement est celui classiquement observé pour ce type de stratégies.

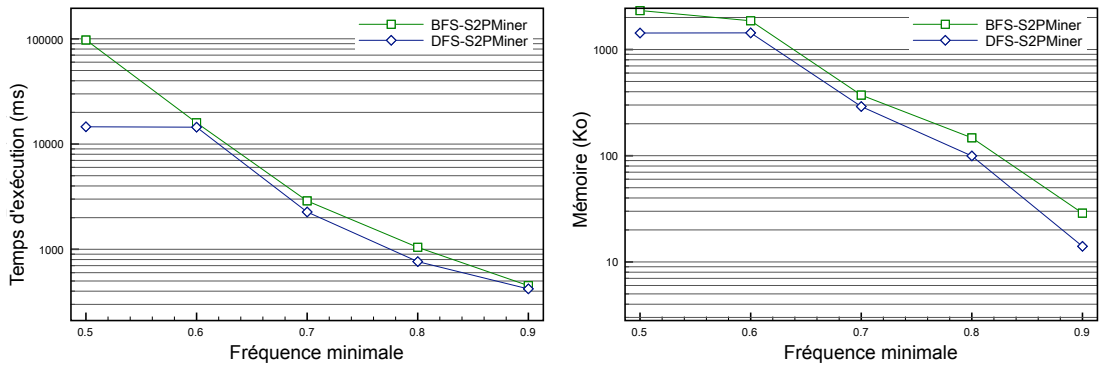


FIGURE 2.11 – Temps d'exécution et mémoire utilisée par l'algorithme par niveau (BFS-S2PMiner) et l'algorithme en profondeur (DFS-S2PMiner) sur un jeu de données synthétiques de type graphe avec 20 zones, 70 dates et 4 zones voisines par zone (Graph20x70)

On remarque que le nombre de voisins sur les performances de l'extraction est très important (cf. figure 2.12). Ce constat met en avant le complexité d'extraire ces motifs par rapport aux motifs séquentiels classiques. Le langage est plus riche avec un espace de recherche plus grand. L'intégration du voisinage spatial implique aussi plus de projections et plus d'informations à stocker en mémoire par l'algorithme.

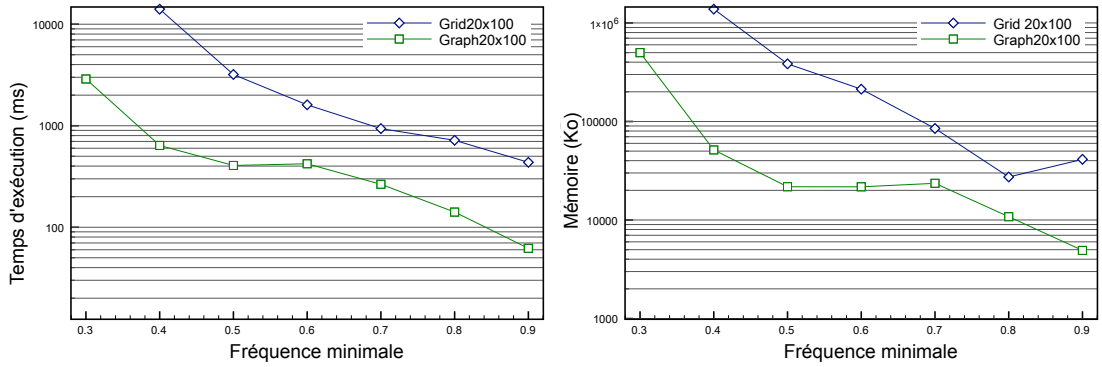


FIGURE 2.12 – Impact du nombre de voisins par zone sur le temps d'exécution et la mémoire utilisée par DFS-S2PMINER ("grid"= 8 voisins et "graph"=4 voisins)

Les résultats expérimentaux ont aussi mis en avant l'impact important du nombre de zones et de dates sur l'extraction, tout comme pour les motifs séquentiels classiques. L'augmentation du nombre de dates semble particulièrement dégrader les performances (plus que le nombre de

zones). Par ailleurs, la contrainte choisie a peu d'influence sur les performances mais elle a un impact très important sur le nombre de motifs extraits (cf. figure 2.13). Dans nos expérimentations, l'indice de participation spatio-temporelle ($STPi$) filtre beaucoup plus de motifs que le support classique.

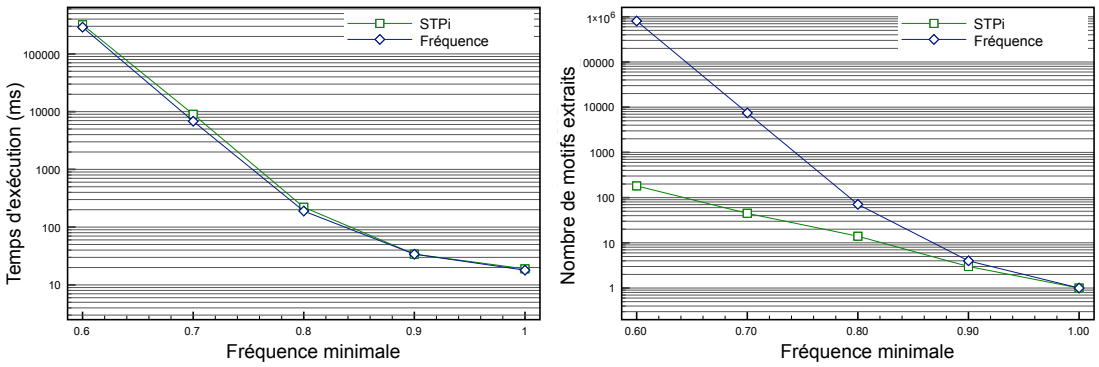


FIGURE 2.13 – Impact de la contrainte (fréquence ou indice de participation) sur le temps d'exécution et le nombre de motifs extraits

3

Recherche d'évolutions fréquentes en utilisant un graphe orienté acyclique attribué

Le travail précédent a abouti à la définition d'un domaine de motifs séquentiels plus riche d'un point de vue spatial. Pour cela, il s'appuie sur une transformation des données en base de données séquentielles, et sur un pré-calcul des relations de voisinage. Toutefois, cette représentation des données reste limitée. Elle ne permet pas de capturer les dynamiques complexes de certains objets spatiaux tels que des accroissements/rétrécissements, des apparitions/disparitions ou des fusions/divisions (comme illustré en figure 3.1). Dans [SFF11], nous avons montré qu'utiliser des approches classiques pour ce type de données aboutissait systématiquement à une perte de données ou à la sur-expression de certains motifs.

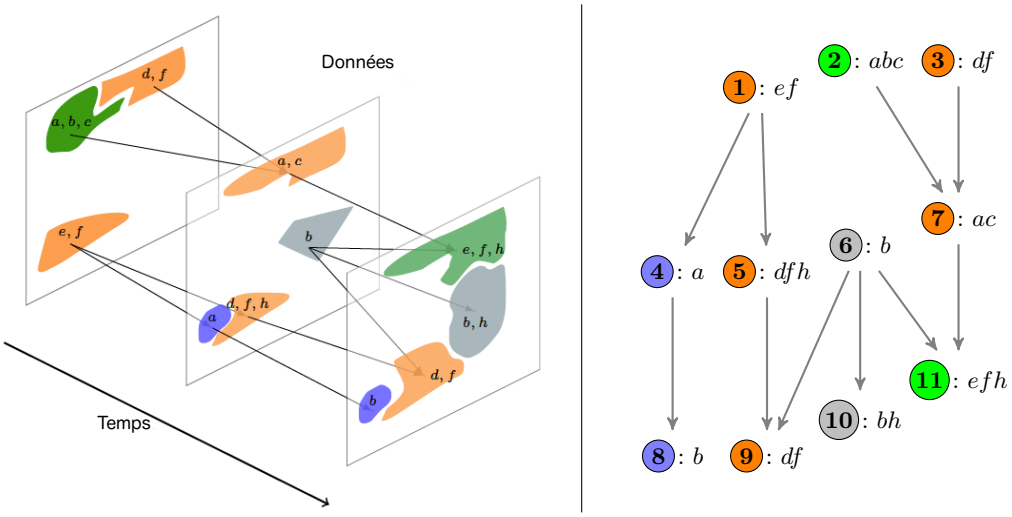


FIGURE 3.1 – Exemple d'objets avec des dynamiques spatiales complexes et leur représentation sous forme de graphe.

Face à ce problème, une solution est d'utiliser une représentation sous forme de graphe. En effet, les graphes permettent de modéliser des phénomènes complexes et variés, ce qui explique l'engouement de la communauté autour de ce type de données. Plus précisément, nous proposons dans ce chapitre de représenter de telles données par un unique graphe orienté acyclique, appelé

a-DAG, où chaque noeud est associé à un ensemble d'attributs. Les motifs recherchés dans ces graphes correspondent à des évolutions fréquentes des caractéristiques associées aux objets spatiaux. Ce domaine de motifs est directement dérivé des motifs séquentiels et des motifs de chemins (*path patterns*), mais leur extraction dans un unique graphe (et non dans un ensemble de graphes) est une tâche beaucoup plus complexe d'un point de vue algorithmique.

Ce travail a été réalisé dans la deuxième partie du projet ANR FOSTER. Bien que les méthodes proposées soient génériques, elles ont donc été principalement appliquées à la problématique de l'érosion des sols en Nouvelle-Calédonie. Les données considérées sont toutefois différentes de celles étudiées au chapitre 1. Nous considérons ici en entrée un ensemble de données hétérogènes sur une région (série d'images satellitaires et base de données géographiques), et non plus une simple base de données d'évènements.

Le tableau 3.1 présente les différents étudiants (stagiaires et doctorant) ayant travaillé sur cette problématique. Il présente aussi les projets de recherche et les collaborations associés à ce travail. Suite à ce tableau sont également listées les principales publications.

Master/Thèse
C. Mu (2014) stage ENSIMAG co-encadrement N. Selmaoui-Folcher (UNC)
M. Collin (2015) stage Université de Bretagne Occidentale co-encadrement N. Selmaoui-Folcher (UNC)
J. Sanhes (2011-2014) thèse dirigée par N. Selmaoui-Folcher (UNC) et J.-F. Boulicaut (INSA Lyon), co-encadrant : F. Flouvat

Projets
projet CNRT "Fonctionnement des petits bassins versants" (2010-2014)
ANR FOSTER (2011-2014) "FOuille de données Spatio-Temporelles : application à la compréhension et à la surveillance de l'ERosion"

Collaborations
INSA Lyon, CNRS, IRD

TABLE 3.1 – Synthèse des encadrements, des projets et des collaborations en lien avec l'extraction de motifs dans un a-DAG

Principales publications
Nazha Selmaoui-Folcher and Frédéric Flouvat. How to use "classical" tree mining algorithms to find complex spatio-temporal patterns ? In Proceedings of the International Conference on Database and Expert Systems Applications (DEXA'11), Toulouse, France, August 2011.
Jérémy Sanhes, Frédéric Flouvat, Claude Pasquier, Nazha Selmaoui-Folcher and Jean-François Boulicaut. Weighted paths as a condensed pattern in a single attributed DAG. In Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI'13), Beijing, China, August 3-9, 2013.
Maxime Collin, Frédéric Flouvat and Nazha Selmaoui-Folcher. PaTSI - Pattern mining of time series of satellite images in KNIME. In Proceedings of the 16th IEEE International Conference on Data Mining (ICDM'16), Demos Session, Barcelona, Spain, December 13-15, 2016.

TABLE 3.2 – Synthèse des publications en lien avec l'extraction de motifs dans un a-DAG

Ce chapitre est organisé de la manière suivante. La section 3.1 présente le domaine de motifs proposé et les contraintes utilisées. La section 3.2 décrit un algorithme dérivé de $PrefixSpan$ pour extraire ces motifs dans un unique graphe orienté acyclique attribué. La section 3.3 présente le processus mis en place pour extraire ces motifs dans une série d'images satellitaires. Pour finir, la section 3.4 discute des résultats obtenus sur différents jeux de données synthétiques et réels, avec un focus plus particulier sur les motifs extraits dans les données liées à l'érosion des sols.

3.1 Cadre théorique

3.1.1 Les données

Dans ce chapitre, nous considérons une région dont l'évolution est décrite par un ensemble d'informations de nature hétérogène : des champs continus (p.ex. type de sol, météorologie, ou élévation), mais aussi des évènements ou des objets plus ponctuels (p.ex. routes, bâtiments ou mines). Ces données sont issues de la fusion d'images satellitaires et de bases de données géographiques. Les modèles spatiaux suivis par ces données sont donc très différents (rasters, points irrégulièrement espacés, courbe de niveau, ou pavage de polygones). Face à cette diversité de modèles, nous avons suivi une approche orientée objets. Les données sont agrégées et discrétisées autour d'objets spatiaux, et l'analyse est conduite sur ces objets. Comme nous les verrons en détail plus tard, ces objets sont identifiés à partir des images satellitaires grâce à des techniques de traitements d'images.

Au final, ces objets et les informations associées sont utilisés pour générer un unique graphe orienté acyclique attribué, appelé encore *a-DAG*. Chaque noeud de ce graphe correspond à un objet spatial. Ce noeud est associé à un ensemble d'informations (d'attributs) issues de la fusion, de l'agrégation et de la discrétisation des données en entrée. Chaque arête orientée représente l'évolution d'un objet entre deux temps. Bien que notre application soit centrée sur le suivi d'objets spatiaux, nous avons travaillé sur une formalisation et une méthode d'extraction génériques afin qu'elles puissent être appliquées à d'autres contextes (p.ex. navigation Web, interactions entre gènes ou dépendances entre code logiciel). Aucune hypothèse n'est donc prise sur les graphes orientés acycliques attribués en entrée.

Plus formellement, un graphe orienté acyclique attribué noté $G = (V_G, E_G, \lambda_G)$ sur un ensemble d'*items* $\mathcal{I}$, aussi appelé **a-DAG**, consiste en un ensemble de sommets V_G, un ensemble d'arcs (orientés) $E_G \subseteq V_G \times V_G$ et une fonction d'étiquetage $\lambda_G : V_G \to \mathcal{P}(\mathcal{I})$ qui associe à chaque sommet du graphe G un sous-ensemble de $\mathcal{I}$ [3]. Un exemple de a-DAG est donné en figure 3.2. Dans le reste de chapitre, nous noterons simplement xy l'itemset $\{x, y\}$ afin d'alléger les notations. S'il existe un chemin du noeud u vers le noeud v dans G, alors u est appelé l'**ancêtre** de v (v est appelé le **descendant**). Si $(u) \rightarrowtail (v) \in E_G$ (i.e. u est un ancêtre immédiat de v), alors u est un **parent** de v (v est un **enfant** du u).

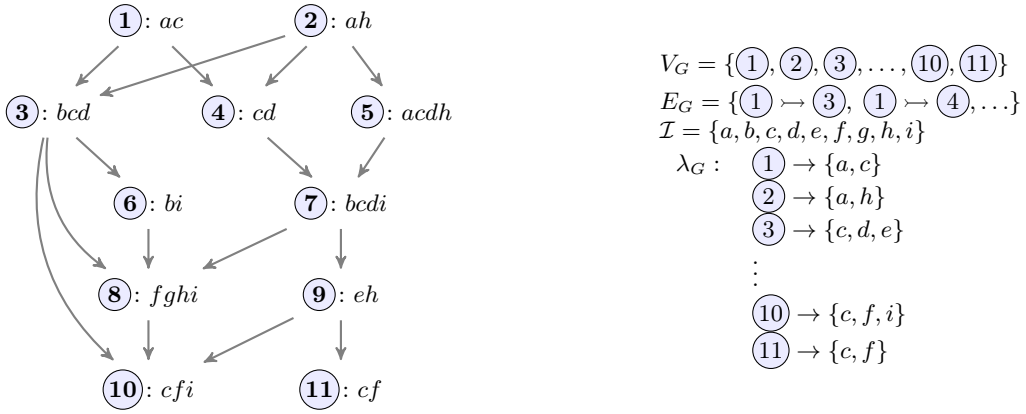


FIGURE 3.2 – Exemple de a-DAG.

3. la notation $\mathcal{P}(\mathcal{I})$ désigne l'ensemble des parties de $\mathcal{I}$

3.1.2 Les chemins pondérés

Dans ce travail, nous avons introduit un nouveau domaine de motifs appelé chemin pondéré. Il s'agit d'un domaine de motifs "hybride" entre les séquences d'*itemsets* et les chemins d'*items* étudiés dans la littérature. Il a l'avantage d'offrir un bon compromis entre expressivité et complexité d'extraction. Dans notre contexte, ce domaine de motifs représente l'évolution d'objets spatiaux.

Un **chemin pondéré** $P = I_1 \overset{w_1}{\rightarrowtail} I_2 \overset{w_2}{\rightarrowtail} \ldots \overset{w_{k-1}}{\rightarrowtail} I_k$ est une séquence d'*itemsets* $I_i \in 2^{\mathcal{I}}$ associés à des noeuds parents-enfants, dans lequel un poids w_i (représentant une fréquence) est associé à chaque transition. Une **occurence** $O = (v_1) \rightarrowtail (v_2) \rightarrowtail (\cdots) \rightarrowtail (v_{|P|})$ du motif P est un séquence de noeuds de G contenant les *itemsets* de P. Dans la figure 3.2, les occurrences du chemin $ah \overset{4}{\rightarrowtail} cd \overset{6}{\rightarrowtail} i$ de taille 3 sont ②↣③↣⑥, ②↣③↣⑧, ②↣③↣⑩, ②↣④↣⑦, ②↣⑤↣⑦, et ⑤↣⑦↣⑧. Les poids 4 et 6 indiquent que la transition $ah \rightarrowtail cd$ apparaît quatre fois, et qu'elle est suivie par six transitions $cd \rightarrowtail i$.

L'intérêt des poids est de pouvoir différencier des motifs qui pourraient paraître similaires avec des domaines de motifs classiques, et ceci sans augmenter la "complexité" du langage de motifs. Par exemple, dans la figure 3.3, le même chemin $a \rightarrowtail b \rightarrowtail c \rightarrowtail d$ dans deux graphes G_1 et G_2 en entrée sera associé à deux motifs différents : $a \overset{1}{\rightarrowtail} b \overset{2}{\rightarrowtail} c \overset{6}{\rightarrowtail} d$ et $a \overset{1}{\rightarrowtail} b \overset{1}{\rightarrowtail} c \overset{1}{\rightarrowtail} d$.

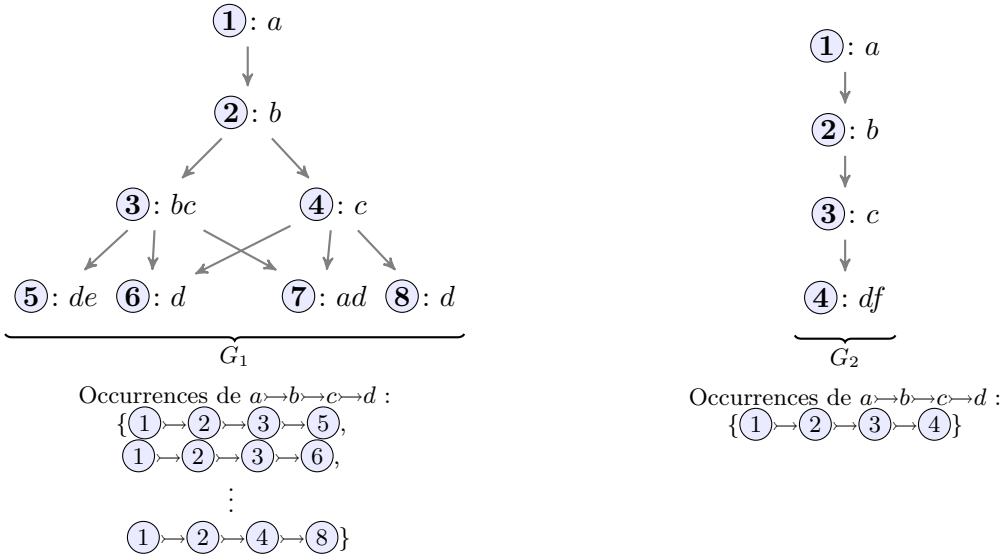


FIGURE 3.3 – a-DAGs dans lesquels le chemin $a \rightarrowtail b \rightarrowtail c \rightarrowtail d$ apparaît de façon très différente.

3.1.3 Les contraintes de fréquence minimale et de non-redondance

Afin de filtrer des motifs pertinents pour les experts, nous introduisons deux contraintes. La première est basée sur la définition de fréquence faite dans [BN08] dans le cadre du graphe unique. La deuxième vise à supprimer les solutions redondantes en adaptant le principe de motifs fermés à notre contexte.

La **fréquence** d'un chemin pondéré $P = I_1 \overset{w_1}{\rightarrowtail} I_2 \overset{w_2}{\rightarrowtail} \ldots \overset{w_{k-1}}{\rightarrowtail} I_k$ dans un a-DAG G est définie comme la fréquence minimale de ses arêtes, i.e.

$$supp(P, G) = \min_{1 \leq i < |P|} w_i = \min_{1 \leq i < |P|} |\{\text{occurrences de } I_i \rightarrowtail I_{i+1} \text{ parmi les occurrences de } P\}|$$

Cette définition garantit l'anti-monotonie de la contrainte de fréquence (et sa capacité d'élagage), et elle est simple à calculer (deux points fondamentaux lorsque l'on veut définir une mesure de fréquence dans le cadre du "graphe unique", cf. section 2.1.2). A partir de cette définition, nous pouvons donc définir la **contrainte de fréquence minimale** $Q_{freq} \equiv supp(P, G) \geq minsup$.

L'ensemble des motifs fréquents peut être très important et contenir beaucoup de redondances. Un motif est considéré comme redondant s'il existe un motif solution plus spécifique associé aux mêmes noeuds. Autrement dit, le motif apparaît exactement aux mêmes endroits aux mêmes moments. Ce concept correspond à la définition de motifs fermés dans les bases de données transactionnelles. Toutefois, cette définition ne tient plus dans le cas d'un graphique unique, car il n'y a pas de transactions et la fermeture d'un motif n'est pas nécessairement unique. Une solution consiste à conserver des motifs "fermés localement", i.e. des motifs maximaux par rapport à l'inclusion représentant des chemins différents dans les données.

Soient deux motifs solutions $P = I_1 \overset{w_1}{\rightarrowtail} I_2 \overset{w_2}{\rightarrowtail} ... \overset{w_{m-1}}{\rightarrowtail} I_m$ et $P' = I'_1 \overset{w'_1}{\rightarrowtail} I'_2 \overset{w'_2}{\rightarrowtail} ... \overset{w'_{n-1}}{\rightarrowtail} I'_n$. P est **redondant** par rapport à P', noté $P \sqsubseteq P'$, si et seulement si $\exists k \in \{1, ..., n\}$ tel que $\forall i \in \{1, ..., m\}$, $I_i \subseteq I'_{k+i-1}$ et $w_i = w'_{k+i-1}$. Notons $Th(G, Q_{freq})$ l'ensemble des chemins pondérés fréquents. La **contrainte de non-redondance** peut donc être définie de la manière suivante

$$Q_{nonRedund} \equiv P \in Th(G, Q_{freq}) \mid \nexists P' \in Th(G, Q_{freq}) \text{ tel que } P \sqsubseteq P'$$

3.1.4 Problématique

Etant donné un unique a-DAG G, l'objectif est d'énumérer l'ensemble des chemins pondérés fréquents non-redondants de G, noté $Th(G, Q_{freq} \wedge Q_{nonRedund})$.

3.2 Stratégie d'extraction des chemins pondérés fréquents

L'extraction des motifs suit globalement une stratégie "*pattern-growth*" similaire à *PrefixSpan* [PHMA+01]. Toutefois, les extensions et les projections effectuées sont spécifiques à notre approche. Chaque extension étend le motif par un *itemset* et non un simple *item*. De plus, la complétude dans notre contexte est plus complexe à obtenir et nécessite un "retour arrière" pour mettre à jour certains préfixes. Une structure de données dédiée a aussi été développée pour effectuer efficacement les projections successives du graphe.

3.2.1 Extensions des motifs et projections du graphe

Une extension de chemin pondéré est exécutée en ajoutant à la fin une arête pondérée et un *itemset*. Elle est effectuée en fonction des occurrences du motif dans le a-DAG en entrée. Les derniers sommets de chaque occurrence, et leurs descendants, sont plus particulièrement déterminants. En d'autres termes, étant donné un motif P, les extensions de P peuvent être trouvées dans la projection du a-DAG par rapport à P. Prenons l'exemple du motif $P = ah \overset{3}{\rightarrowtail} cd \overset{3}{\rightarrowtail} bi$ dans le a-DAG G de la figure 3.2. Ce motif est associé aux occurrences $\{②\rightarrowtail③\rightarrowtail⑥, ②\rightarrowtail④\rightarrowtail⑦, ②\rightarrowtail⑤\rightarrowtail⑦\}$. Comme illustré dans la figure 3.4, la projection de G par rapport P (figure de droite) représente toutes les extensions possibles de ce motif.

Pour éviter la construction de motifs redondants, l'extension doit être effectuée de telle sorte qu'il n'y ait pas d'autre extension possible avec le même poids et le même préfixe. En d'autres termes, les motifs résultants doivent être maximaux et associés à des occurrences différentes

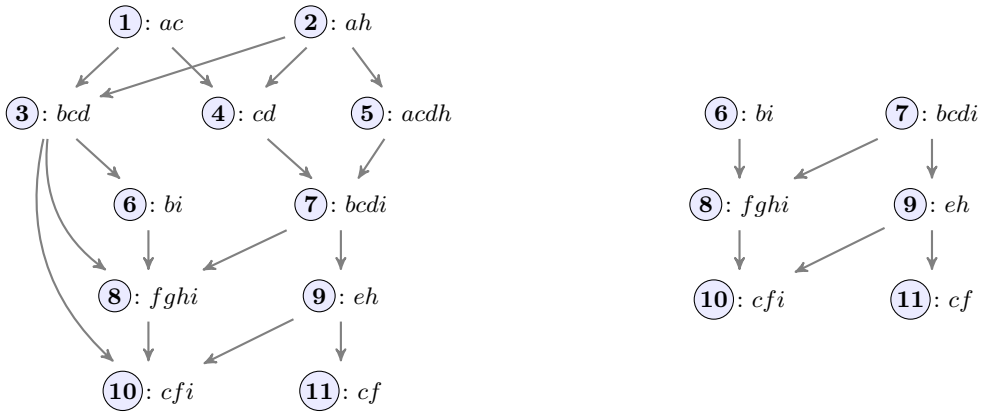


FIGURE 3.4 – Exemple de a-DAG G (gauche) et de sa projection par rapport à P (droite), avec $P = cd \overset{3}{\rightarrowtail} bi$.

dans le a-DAG. La figure 3.5 illustre toutes les extensions maximales (en termes d'*itemsets*) du chemin pondéré $ah \overset{3}{\rightarrowtail} cd \overset{3}{\rightarrowtail} bi$. Par exemple, le motif peut être étendu avec $\overset{3}{\rightarrowtail} h$ relativement aux arêtes $\{(6) \rightarrowtail (8), (7) \rightarrowtail (8), (7) \rightarrowtail (9)\}$ et $\overset{2}{\rightarrowtail} fghi$ relativement aux arêtes $\{(6) \rightarrowtail (8), (7) \rightarrowtail (8)\}$. Notez que seule la première extension est fréquente avec $minsup = 3$. Ces extensions sont appelées "**extensions complètes**" car toutes les occurrences du motif peuvent être étendues par une arête et un noeud associé au même *itemset* dans le a-DAG en entrée.

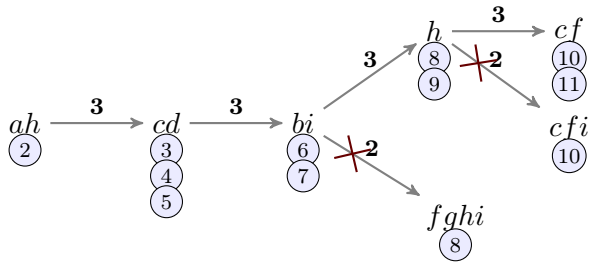
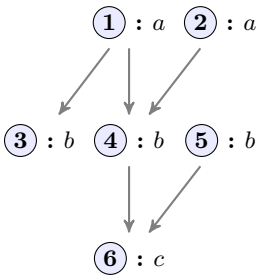


FIGURE 3.5 – Extensions "complètes" de $ah \overset{3}{\rightarrowtail} cd \overset{3}{\rightarrowtail} bi$ (minsup=3).

L'avantage des extensions complètes est de ne pas impacter le préfixe du motif. Par conséquent, si le motif était initialement fréquent et non redondant, le motif étendu le reste, ce qui simplifie beaucoup les tests. Cette propriété est celle exploitée par $PrefixSpan$. Toutefois, cette approche seule peut manquer des motifs dans notre contexte. Par exemple, il y a trois solutions dans la figure 3.6 : $a \overset{3}{\rightarrowtail} b$, $b \overset{2}{\rightarrowtail} c$ et $a \overset{2}{\rightarrowtail} b \overset{1}{\rightarrowtail} c$. Les deux premières peuvent être générées en utilisant une extension complète des motifs a et b, mais la dernière ne le peut pas car c n'est pas associé à un noeud fils du noeud 3. Or, les occurrences de $a \overset{3}{\rightarrowtail} b$ sont $\{(1) \rightarrowtail (3), (1) \rightarrowtail (4), (2) \rightarrowtail (4)\}$.

Pour trouver ces motifs, nous devons considérer des **extensions partielles**, c'est-à-dire des extensions qui n'étendent que certaines occurrences du motif. Toutefois, ces extensions sont plus difficiles à traiter car elles peuvent impacter les poids et/ou les *itemsets* du préfixe. Dans l'exemple précédent, $a \overset{3}{\rightarrowtail} b$ peut être partiellement étendu avec $\overset{2}{\rightarrowtail} c$, mais cela élimine



Chemins pondérés fréquents non redondants (minsup=1) :
$a \overset{3}{\rightarrowtail} b$
$b \overset{2}{\rightarrowtail} c$
$a \overset{2}{\rightarrowtail} b \overset{1}{\rightarrowtail} c$
Chemins pondérés fréquents non redondants générés par une extension complète :
$a \overset{3}{\rightarrowtail} b$
$b \overset{2}{\rightarrowtail} c$
Chemins pondérés fréquents non redondants générés par une extension partielle :
$a \overset{2}{\rightarrowtail} b \overset{1}{\rightarrowtail} c$

FIGURE 3.6 – Exemple de a-DAG illustrant l'importance des extensions partielles.

l'occurrence $\textcircled{1} \rightarrowtail \textcircled{3}$ de $a \overset{3}{\rightarrowtail} b$. Ainsi, le poids de $a \rightarrowtail b$ doit être modifié et nous obtenons le motif $a \overset{2}{\rightarrowtail} b \overset{1}{\rightarrowtail} c$. La figure 3.7 illustre un autre exemple d'extension partielle pour le motif $ah \overset{3}{\rightarrowtail} cd \overset{3}{\rightarrowtail} bi$ issu du a-DAG de la figure 3.2. Ce motif peut être partiellement étendu avec $\overset{1}{\rightarrowtail} eh$. Au delà d'être non fréquente, cette extension a aussi un impact sur le préfixe du motif, car eh est seulement associé avec un noeud fils de $\textcircled{7}$. L'occurrence $\textcircled{2} \rightarrowtail \textcircled{3} \rightarrowtail \textcircled{6}$ n'est donc plus valide, et le motif étendu devient $ah \overset{2}{\rightarrowtail} cd \overset{2}{\rightarrowtail} bi \overset{1}{\rightarrowtail} eh$, avec pour occurrences $\{\textcircled{2} \rightarrowtail \textcircled{4} \rightarrowtail \textcircled{7} \rightarrowtail \textcircled{9}, \textcircled{2} \rightarrowtail \textcircled{5} \rightarrowtail \textcircled{7} \rightarrowtail \textcircled{9}\}$. Toutefois, ce motif n'est pas maximal par rapport à ces occurrences de chemins. Ce motif est donc redondant. En effet, $ah \overset{2}{\rightarrowtail} cd \overset{2}{\rightarrowtail} bcdi \overset{1}{\rightarrowtail} eh$ est associé aux mêmes occurrences et il inclut le précédent.

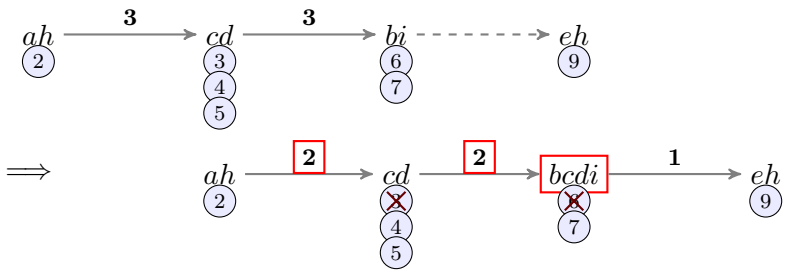


FIGURE 3.7 – Extensions partielles de $ah \overset{3}{\rightarrowtail} cd \overset{3}{\rightarrowtail} bi$.

Une extension partielle peut conduire à des solutions non générées par des extensions complètes, mais il peut aussi conduire à un motif redondant ou non fréquent. Après une extension complète, il suffit de tester l'extension pour savoir si le motif généré est une solution. Après une extension partielle, il faut tester l'extension du motif, mais aussi son préfixe. Des optimisations peuvent être mises en place pour limiter ce surcoût. Par exemple, il est possible de calculer directement la fréquence en identifiant les occurrences qui ne sont plus valides. De même, il est possible de générer le préfixe maximal en faisant l'intersection des *itemsets* associés aux occurrences.

3.2.2 Parcours en profondeur, structure de données et optimisations

La stratégie d'exploration utilisée pour extraire ces motifs s'appuie sur un parcours en profondeur de l'espace de recherche. A chaque étape, l'algorithme se concentre sur une projection du a-DAG. Ensuite, il utilise des extensions complètes et partielles pour générer des motifs fréquents

non redondants liés à cette projection.

Une approche basique pour trouver toutes les solutions consisterait à construire de manière récursive une projection complète du a-DAG pour chaque préfixe. Toutefois, le coût en mémoire de cette approche serait très élevé. Pour résoudre ce problème, nous ne gardons pas toutes les arêtes de l'a-DAG projeté, mais seulement celles nécessaires pour les premières extensions du motif en cours d'étude, également appelées arêtes candidates. Par exemple, les arêtes candidates de $P = ah \overset{3}{\rightarrowtail} cd \overset{3}{\rightarrowtail} bi$ (figure 3.4) sont $\{(6,8),(7,8),(7,9)\}$. Ces arêtes permettent de générer toutes les extensions de taille un de P (i.e. $P \overset{3}{\rightarrowtail} h$, $P \overset{2}{\rightarrowtail} fghi$ et $P \overset{1}{\rightarrowtail} eh$).

Cet exemple montre aussi que l'extension des motifs peut être transformée en problème d'extraction d'*itemsets* fermés fréquents dans cet ensemble d'arêtes candidates. La base de données transactionnelle serait $\{fghi, fghi, eh\}$ car il y a deux arêtes conduisant à $fghi$ et une pour eh. Les *itemsets* fermés seraient h (fréquence : 3), eh (fréquence : 1) et $fghi$ (fréquence : 2). Ces trois *itemsets* fermés permettent de générer les trois extensions non redondantes possibles de P. Sur la base de cette transformation, l'algorithme génère de manière récursive chaque motif à partir des *itemsets* fermés fréquents extraits des arêtes candidates. Si un *itemset* aboutit à une extension partielle, l'algorithme met d'abord à jour les poids du préfixe. Ensuite, si le motif résultant est fréquent, les *itemsets* du préfixe sont recalculés de façon à être maximaux (en faisant des intersections des *itemsets* associés aux noeuds). Après, l'algorithme continue ses extensions récursives, tout en enregistrant au fur et à mesure les solutions. La figure 3.8 présente une partie de la trace de cet algorithme sur le a-DAG utilisé précédemment.

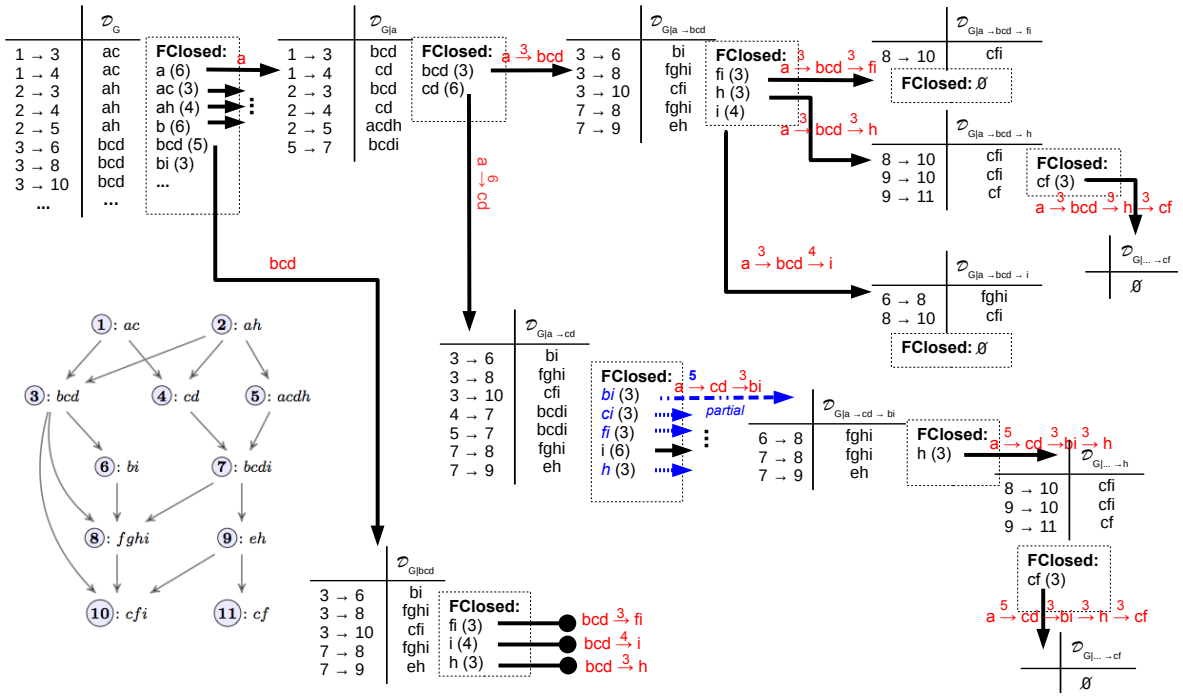


FIGURE 3.8 – Exemple d'exécution de l'algorithme avec $minsup = 3$.

Tous les motifs fréquents non redondants sont générés par cet algorithme en profondeur. Toutefois, certains fragments de motifs peuvent être générés plusieurs fois. Par exemple, dans la figure 3.8, $a \overset{3}{\rightarrowtail} bcd \overset{3}{\rightarrowtail} fi$ est généré à partir de l'extension du préfixe a, et $bcd \overset{3}{\rightarrowtail} fi$

est généré à partir de l'extension du préfixe *bcd*. Pour éviter cela, un test d'inclusion doit être effectué (et certaines solutions mises à jour) avant d'étendre récursivement le motif courant. En effet, si $a \overset{3}{\rightarrowtail} bcd \overset{3}{\rightarrowtail} fi$ a été généré en premier, l'extension de $bcd \overset{3}{\rightarrowtail} fi$ doit être arrêtée et ce dernier motif ne doit pas être enregistré. Si $bcd \overset{3}{\rightarrowtail} fi$ a été généré en premier, ce motif doit être mis à jour avec le préfixe $a \overset{3}{\rightarrowtail}$ et son extension doit être arrêtée (car déjà générée lors de l'extension de $bcd \overset{3}{\rightarrowtail} fi$).

Cette approche générale peut toutefois ne pas passer à l'échelle sans une structure de données adaptée. Trois opérations sont particulièrement coûteuses dans notre cas : 1) la mise à jour du préfixe suite à une extension partielle ; 2) le test d'inclusion évitant de générer plusieurs fois les mêmes fragments de motif ; 3) la mise à jour du préfixe si ses extensions ont déjà été explorées. Face à ces opérations, une structure de données, appelée **graphe de motifs**, a été proposée afin d'extraire les solutions plus efficacement. Chaque chemin de ce graphe représente un motif solution et ses occurrences. Ce graphe, noté G_{Th}, est composé d'un ensemble de noeuds V_{Th}, d'un ensemble d'arêtes $E_{Th} \subseteq V_{Th} \times V_{Th}$, d'une fonction d'étiquetage $\lambda_w : E_{Th} \rightarrow \mathbb{Z}$ qui associe chaque arête de E_{Th} à un poids, et d'une autre fonction d'étiquetage $\lambda_{Th} : V_{Th} \rightarrow (2^{\mathcal{I}}, 2^{V_G})$ qui associe chaque noeud de V_{Th} à un *itemset* et à un sous-ensemble de noeuds du a-DAG en entrée G. La figure 3.9 présente le graphe de motifs obtenu à partir des cinq solutions extraites dans la figure 3.8. Par exemple, le chemin $\textcircled{1} \rightarrowtail \textcircled{2} \rightarrowtail \textcircled{3}$ représente le motif $a \overset{3}{\rightarrowtail} bcd \overset{3}{\rightarrowtail} fi$ et ses occurrences $\{\textcircled{1} \rightarrowtail \textcircled{3} \rightarrowtail \textcircled{8}, \textcircled{1} \rightarrowtail \textcircled{3} \rightarrowtail \textcircled{10}, \textcircled{2} \rightarrowtail \textcircled{3} \rightarrowtail \textcircled{8}, \textcircled{2} \rightarrowtail \textcircled{3} \rightarrowtail \textcircled{10}, \textcircled{5} \rightarrowtail \textcircled{7} \rightarrowtail \textcircled{8}\}$. Comme illustré par cette figure, un noeud de ce graphe peut être partagé entre plusieurs motifs solutions.

Exemples de chemins fréquents non redondants (minsup=3) :
- $a \overset{3}{\rightarrowtail} bcd \overset{3}{\rightarrowtail} fi$
- $a \overset{3}{\rightarrowtail} bcd \overset{3}{\rightarrowtail} h \overset{3}{\rightarrowtail} cf$
- $a \overset{3}{\rightarrowtail} bcd \overset{4}{\rightarrowtail} i$
- $a \overset{6}{\rightarrowtail} cd$
- $a \overset{5}{\rightarrowtail} cd \overset{3}{\rightarrowtail} bi \overset{3}{\rightarrowtail} h \overset{3}{\rightarrowtail} cf$

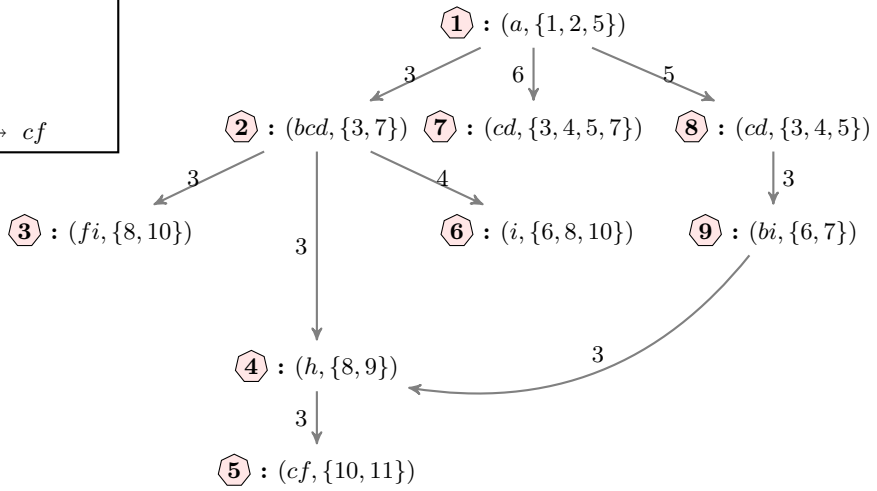


FIGURE 3.9 – Graphe de motifs des solutions extraites dans la figure 3.8.

Cette structure de données permet de stocker efficacement les motifs, mais aussi de les générer efficacement. Chaque extension d'un motif correspond à la génération d'un nouveau noeud et d'une nouvelle arête dans le graphe de motifs. Si le noeud est déjà dans le graphe de motifs, une simple arête est ajoutée et les extensions sont arrêtées pour ce préfixe (car déjà traitées dans une itération précédente). Nous avons un tel exemple dans la figure 3.9 lors de l'extension du préfixe

$a \overset{5}{\rightarrowtail} cd \overset{3}{\rightarrowtail} bi$ (chemin $\langle 1 \rangle \rightarrowtail \langle 8 \rangle \rightarrowtail \langle 9 \rangle$) avec $\overset{3}{\rightarrowtail} h \overset{3}{\rightarrowtail} cf$ (chemin $\langle 4 \rangle \rightarrowtail \langle 5 \rangle$). Grâce à cette structure de données, les tests d'inclusion et les mises à jour des préfixes sont beaucoup plus faciles et consistent simplement à rechercher un noeud dans un ensemble de noeuds.

L'algorithme 3 décrit l'utilisation de cette structure de données pour extraire les chemins pondérés fréquents non redondants. Soit un motif $P = I_1 \overset{w_1}{\rightarrowtail} I_2 \overset{w_2}{\rightarrowtail} ... \overset{w_{k-1}}{\rightarrowtail} I_k$. L'algorithme étend ce motif avec toutes les extensions fréquentes non redondantes possibles X. Il génère donc des motifs $P' = I'_1 \overset{w'_1}{\rightarrowtail} I'_2 \overset{w'_2}{\rightarrowtail} ... \overset{w'_{k-1}}{\rightarrowtail} I'_k \overset{supp_X}{\rightarrowtail} X$, avec $I'_i = I_i$ et $w'_i = w_i$, s'il s'agit d'une extension complète, ou $I_i \subseteq I'_i$ and $w'_i \leq w_i$, s'il s'agit d'une extension partielle ($i = 1, ..., k$). Les lignes 5-9 correspondent à la génération du premier *itemset* I_1 de P, i.e. la génération du noeud (I_1, V_1) dans le graphe de motifs. La ligne 6 vérifie si le noeud est déjà dans le graphe de motifs, i.e. si un sur-motif a déjà été généré. Les lignes 11-25 représentent le cas général où I_k est le dernier *itemset* de P et V_k ses occurrences. La ligne 11 construit V'_k, i.e. le sous-ensemble de noeuds de V_k qui peuvent être étendus avec X. Si $V_k \neq V'_k$ (ligne 12), alors il s'agit d'une extension partielle de P avec X. La fonction *BackwardUpdate* en ligne 13 calcule le nouveau préfixe $I'_1 \overset{w'_1}{\rightarrowtail} I'_2 \overset{w'_2}{\rightarrowtail} ... \overset{w'_{k-1}}{\rightarrowtail} I'_k$ à partir de V'_k et retourne "vrai" s'il est fréquent. Le principe de cette fonction est d'explorer de manière récursive tous les ancêtres de u_k et de mettre à jour leurs occurrences en fonction de V'_k (mises à jour en cascade). Cette exploration se poursuit tant que le préfixe reste fréquent. A la fin, si $I'_1 \overset{w'_1}{\rightarrowtail} I'_2 \overset{w'_2}{\rightarrowtail} ... \overset{w'_{k-1}}{\rightarrowtail} I'_k$ est fréquent, l'algorithme insère tous les sommets et arêtes nécessaires dans le graphe de motifs. Les lignes 17-24 représentent

Algorithm 3 PrefixPathGrowth($G, minsup, G_{Th}, u_k$)

Require: An a-DAG $G = (V_G, E_G, \lambda_G)$, a minimum support threshold *minsup*, a pattern graph $G_{Th} = (V_{Th}, E_{Th}, \lambda_{Th}, \lambda_w)$, a vertex $u_k \in V_{Th}$ with $\lambda_{Th}(u_k) = (I_k, V_k)$;

1: $cand(E|_P) = \{(u, v) \in E_G \mid u \in V_k\}$
2: $FClosed = MiningFreqClosedItemset\,(\ cand(E|_P), minsup\ )$
3: **for all** $X \in FClosed$ **do**
4: $V_X = \{v \in V_G \mid (u, v) \in cand(E|_P)$ and $X \subseteq \lambda_G(v)\}$
5: **if** $I_k == \emptyset$ **then**
6: **if** $\nexists u_X \in V_{Th}$ s.t. $\lambda_{Th}(u_X) = (X, V_X)$ **then**
7: Insert u_X in V_{Th} with $\lambda_{Th}\,(u_X) = (X, V_X)$
8: PrefixPathGrowth(G, $minsup$, G_{Th}, u_X)
9: **end if**
10: **else**
11: $V'_k = \{u \in V_k \mid (u, v) \in cand(E|_P)$ and $X \subseteq \lambda_G(v)\}$
12: **if** $V_k \neq V'_k$ **then** // ***Extension partielle***
13: **if** $BackwardUpdate(\ G, minsup, G_{Th}, u_k, V'_k\ ) ==$ False **then**
14: Continue // ***arrêter l'itération courante et passer à l'extension suivante***
15: **end if**
16: **end if**
17: Let $u'_k \in V_{Th}$ s.t. $\lambda_{Th}(u'_k) = (I'_k, V'_k)$
18: **if** $\nexists u_X \in V_{Th}$ s.t. $\lambda_{Th}(u_X) = (X, V_X)$ **then**
19: Insert u_X in V_{Th} with $\lambda_{Th}\,(u_X) = (X, V_X)$
20: Insert (u'_k, u_X) in E_{Th} with $\lambda_w\,((u'_k, u_X)) = support(X)$
21: PrefixPathGrowth(G, $minsup$, G_{Th}, u_X)
22: **else**
23: Insert (u'_k, u_X) in E_{Th} with $\lambda_w\,((u'_k, u_X)) = support(X)$
24: **end if**
25: **end if**
26: **end for**

l'extension du graphe de motifs avec X. u'_k est le noeud courant dans le graphe de motif. Si l'extension est complète, $u'_k = u_k$. Si l'extension est partielle, u'_k est le noeud du graphe de motif associé à V'_k (un sommet existant ou un nouveau créé par la fonction *BackwardUpdate*). La ligne 18 teste si (X, V_X) est déjà dans la structure de données, i.e. si cette partie du motif (et ses extensions) a déjà été générée. Si ce n'est pas le cas, l'algorithme insère le noeud u_X, génère une arête (u'_k, u_X) et continue l'extension de P'. Si ce noeud est déjà dans le graphe de motifs, l'algorithme ne génère qu'une arête (u'_k, u_X) et arrête les extensions.

3.3 Intégration dans un processus d'analyse d'une série d'images satellitaires

L'algorithme d'extraction décrit dans la section précédente a été intégré dans un processus complet d'analyse de séries temporelles de données hétérogènes (rasters et vecteurs). Comme le montre la figure 3.10, ce processus prend en entrée une série d'images satellitaires complétée par des données vectorielles issues de bases de données géographiques (SIG), et extrait les chemins pondérés représentant des évolutions spatio-temporelles fréquentes.

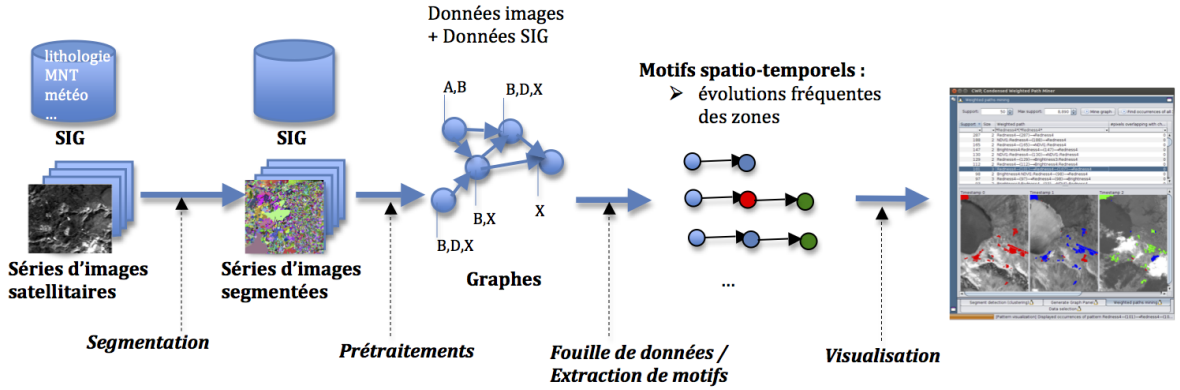


FIGURE 3.10 – Le processus d'analyse d'une série d'images satellitaires

La première étape de ce processus est la segmentation. Pour chaque image satellitaire, la méthode de segmentation par ligne de partage des eaux (*watershed*) [BM93] est utilisée pour regrouper les pixels proches partageant des valeurs radiométriques similaires. Cette méthode classique a été choisie parce qu'elle génère beaucoup de petites régions très homogènes ("sur-segmentation"). Cette méthode a été initialement développée pour traiter des images à bande unique. Cependant, les images satellitaires sont généralement des images multispectrales (i.e. composées de plusieurs bandes). Pour obtenir un résultat de segmentation unique pour chaque image, chaque bande spectrale est traitée indépendamment, puis les régions sont fusionnées en faisant des intersections. Au cours de ce processus, les données vectorielles disponibles sur la zone étudiée sont aussi intégrées. Cette approche de segmentation est décrite dans la figure 3.11. A la fin, chaque image est associée à un ensemble de régions/objets caractérisés par un ensemble de valeurs (moyennes des valeurs radiométriques de l'image et des valeurs catégorielles des données vectorielles). La figure 3.12 montre un exemple de segmentation obtenue sur une image satellitaire.

La deuxième étape du processus est le prétraitement de la série temporelle d'objets générés

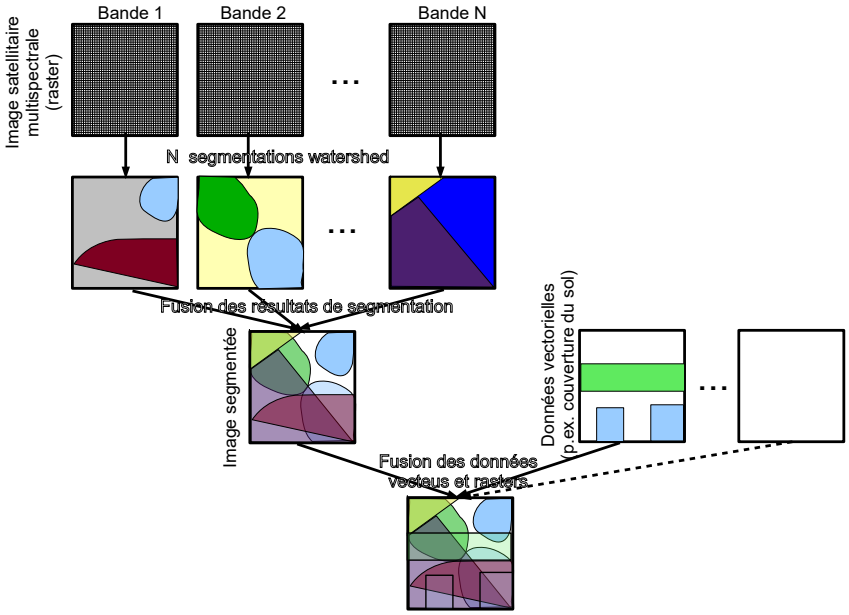


FIGURE 3.11 – Le sous-processus de segmentation

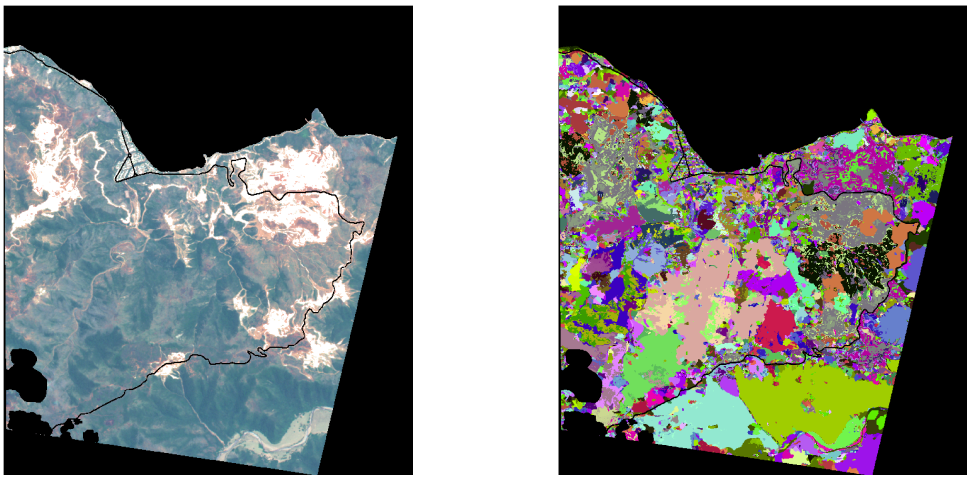


FIGURE 3.12 – Exemple de résultat de segmentation d'une image satellitaire

par l'étape de segmentation. Elle comprend le calcul des indices de télédétection liés à l'érosion des sols, la discrétisation des attributs numériques et la génération du a-DAG.
Dans notre contexte, trois indices de télédétection liés à l'érosion du sol ont été calculés : $NDVI$ (indice de végétation), RI (indice de rougeur) et BI (indice de luminosité). Le $NDVI$ est une mesure commune pour observer la végétation. Il mesure la fraction de rayonnement photosynthétiquement absorbé. Un $NDVI$ faible représente une activité photosynthétique faible, i.e. peu de végétation ou une végétation en mauvaise santé. Le RI est utilisé pour quantifier les sols nus. Il mesure la rougeur des sols. Dans les zones étudiées, les sols sont riches en fer et sont donc rouges rouille. Le BI mesure la luminosité du sol. Il s'agit aussi une mesure commune lors de l'étude des sols érodés [BMBH95].
Ensuite, les attributs numériques sont discrétisés. Dans nos expérimentations, les indices de télédétection de chaque image ont été par exemple discrétisés en cinq intervalles, soit cinq niveaux de 0 à 4. Par exemple, nous avons $NDVI0$, $NDVI1$, $NDVI2$,... $NDVI4$. Ainsi, la discrétisation peut être différente d'une image à l'autre. L'intérêt de cette approche est de faire face aux différentes conditions d'acquisition des images satellitaires (p. ex. étalonnage, capteur, angle, lumière et décalage de valeur).
Pour finir, le a-DAG est construit. Les noeuds du a-DAG sont les différents objets spatiaux/régions des images segmentées. Chaque noeud est étiqueté par un ensemble d'attributs correspondant aux indices de télédétection discrétisés et aux attributs des données vectorielles (p.ex., la couverture terrestre, le type de sol). Les arêtes du a-DAG représentent l'influence/l'évolution possible d'un objet au temps t en objets au temps $t+1$. Deux objets sont reliés par une arête lorsque leur distance est faible et lorsqu'ils apparaissent à deux temps consécutifs. Dans nos expérimentations, deux zones sont liées si leur intersection est supérieure à un seuil donné. Par exemple, un objet o_1 au temps t est lié à un objet o_2 au temps $t+1$ si au moins 10% des pixels de o_1 sont communs avec o_2. La figure 3.1 illustre la génération d'un a-DAG sur un petit exemple.

La troisième étape du processus est l'extraction des chemins pondérés fréquents non redondants à l'aide de l'algorithme en profondeur décrit dans la section précédente.

La dernière étape du processus est la visualisation des motifs extraits. Comme discuté précédemment, le concept de motif est difficile à comprendre pour les experts (géographes ou géologues). De plus, ce domaine de motif est plus complexe que les co-localisations étudiées précédemment. Ainsi, nous avons encore opté pour une approche de visualisation plus classique dans ce travail. L'expert sélectionne un motif dans une liste de solutions et ses occurrences sont affichées sur la série d'images satellitaires en entrée. L'avantage de cette approche est de remettre le motif dans son contexte géographique. Les experts peuvent visualiser où, quand et comment le chemin pondéré fréquent non-redondant (i.e. l'évolution fréquente) se produit. Puisque les occurrences de motifs peuvent commencer à des temps différents, différentes couleurs sont utilisées sur l'image pour identifier la position temporelle des différents objets par rapport au motif. Par exemple, le motif $Redness4 \overset{1405}{\rightarrowtail} Redness4 \overset{1408}{\rightarrowtail} Brightness4$ est affiché en figure 3.13. Les zones rouges correspondent au premier $Redness4$, les zones bleues sont associées au second $Redness4$, et les zones vertes sont associées à $Brightness4$.

3.4 Expérimentations et application à l'étude de l'érosion des sols

3.4.1 Prototype

Le processus d'analyse d'une série d'images satellitaires précédent a été intégré sous forme de *plugin* dans la plate-forme *KNIME* [BCD+07]. Cette plate-forme est un logiciel *open-source* et

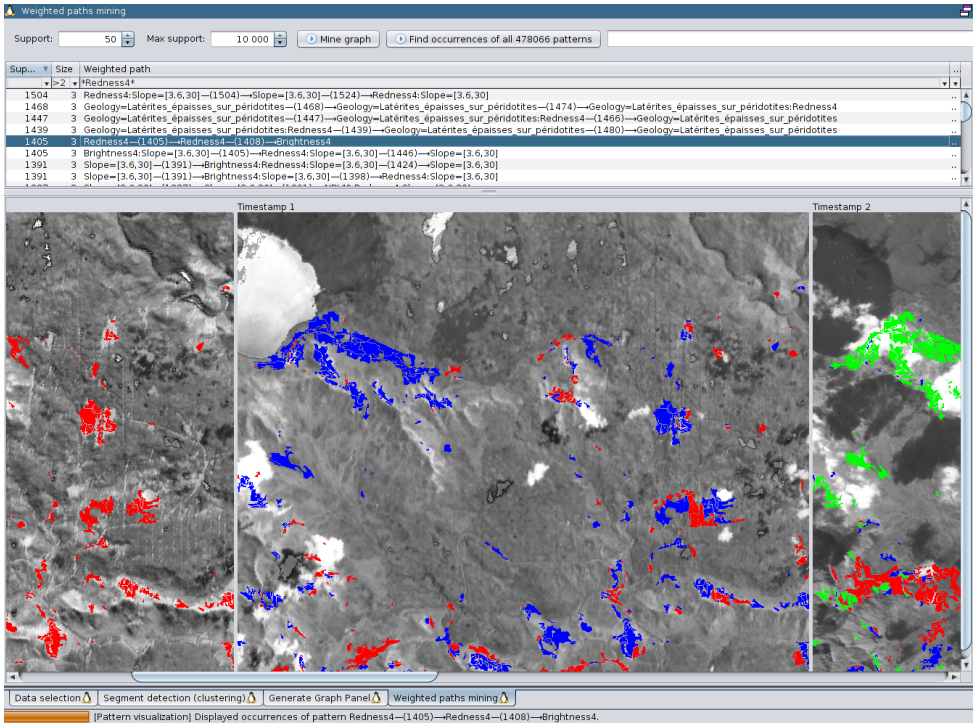


FIGURE 3.13 – Exemple d'affichage d'un motif

libre d'analyse, de visualisation et d'intégration de données. L'intérêt de *KNIME* est de fournir une interface conviviale pour composer des processus *KDD* complexes (*workflows*). Chaque traitement est représenté par un "noeud" et les noeuds sont reliés entre eux via leur(s) entrée(s) et leur(s) sortie(s). Il a une communauté dynamique et intègre déjà un nombre important de traitements issus notamment de l'apprentissage et de la fouille de données. En outre, il fournit un kit de développement logiciel basé sur *Eclipse* pour implémenter de nouveaux noeuds et *plugins* dans le langage de programmation Java.

Le *plugin* que nous avons développé est composé de plusieurs noeuds représentant chacun une étape différente du processus (cf. figure 3.14). Pour garantir le passage à l'échelle, nous n'avons pas entièrement intégré toutes les étapes en tant que noeud Java *KNIME*. Le traitement des images satellitaires nécessite beaucoup de ressources (mémoire et processeur), tout comme l'extraction des chemins pondérés. Ils requièrent des implémentations optimisées et l'utilisation de bibliothèques spécifiques telles que la boîte à outils *Orfeo* [CI09] (une bibliothèque C++ *open-source* du CNES particulièrement optimisée pour le traitement d'images satellitaires). En conséquence, ces étapes ont été implémentées en C++ et profondément optimisées pour traiter des images satellitaires à très haute résolution. Pour l'utilisateur, cette organisation du code est transparente. L'utilisateur manipule les noeuds KNIME. Ces noeuds sont en Java et appellent les implémentations C++ optimisées. Seuls les scripts d'installation doivent être lancés pour compiler le code C++ et installer les bibliothèques. Les données intermédiaires sont transférées de manière transparente d'un noeud à un autre via des fichiers. L'inconvénient de ces fichiers est qu'ils ralentissent un peu l'ensemble du processus. Leur avantage est qu'ils gardent en mémoire les résultats intermédiaires du processus.

Les noeuds *KNIME* développés pour chaque étape ou sous-étape sont indépendants. Ils sont donc utilisables dans d'autres contextes ou d'autres cas d'utilisation. En particulier, le noeud de

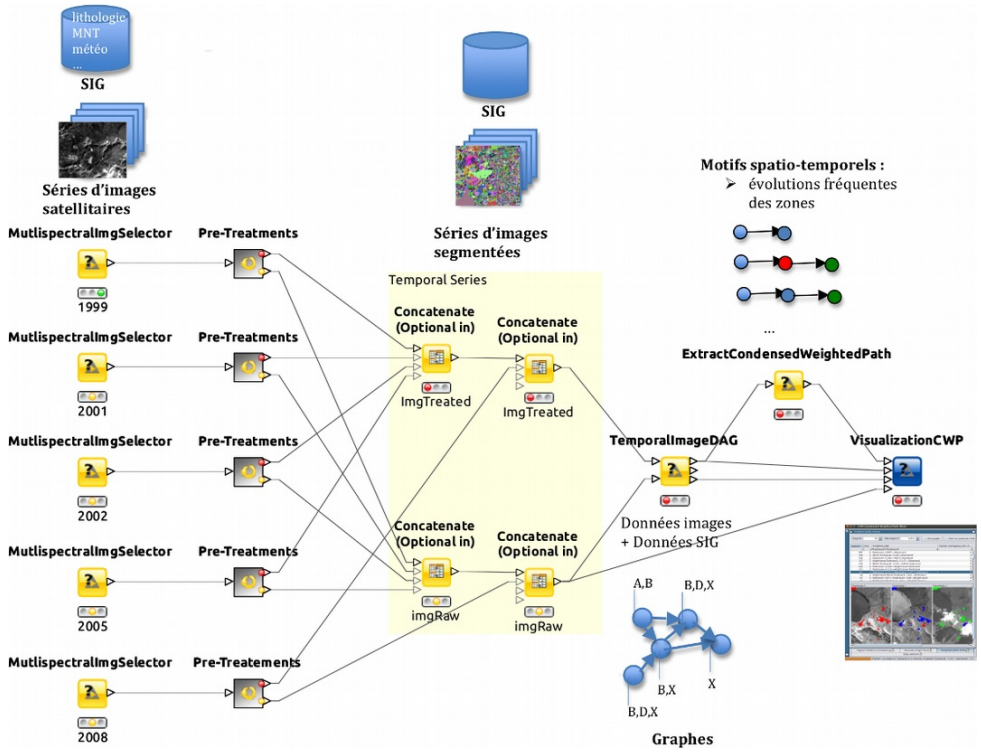


FIGURE 3.14 – Exemple d'utilisation du plugin développé dans *KNIME*

fouille de graphe est générique et peut être appliqué à d'autres problèmes. Ce module n'a besoin que d'un a-DAG en entrée. Toutefois, les noeuds permettant de générer le graphe et de visualiser les motifs solutions sont relativement spécifiques aux séries temporelles d'images satellitaires.

3.4.2 Protocole expérimental

Tout d'abord, nous avons utilisé notre approche pour étudier des données liées au suivi de l'érosion des sols. Ensuite, nous avons démontré la généricité de celle-ci et son passage à l'échelle en utilisant des données issues d'un réseau de citations de brevets ainsi que sur neuf jeux de données synthétiques. L'analyse des performances a étudié le temps d'exécution, la quantité de mémoire utilisée et le nombre de solutions pour différents seuils minimums de fréquence. Le tableau 3.3 présente les caractéristiques des différents jeux de données.

Nous avons notamment étudié une série temporelle d'images satellitaires d'une région touchée par l'érosion du sol. Cette série temporelle est composée de cinq images (SPOT4 et SPOT5) datées en 1999, 2002, 2005, 2008 et 2009. La résolution spatiale de ces images est de 10 mètres. La taille de la zone étudiée est 794×660 pixels, soit 52.5 km^2. En plus de ces données rasters, nous avons aussi les données vectorielles suivantes : une carte des types de sol, des données de couverture terrestre et un modèle d'élévation numérique à partir duquel nous avons dérivé la pente. Comme indiqué précédemment, les indices $NDVI$, RI et BI ont été calculés, et discrétisés en cinq classes d'intensité. Les objets ont été identifiés par segmentation puis transformés en a-DAG. Deux objets sont liés par une arête s'ils partagent au moins 10% de leurs pixels et qu'ils sont à deux temps consécutifs.

Le réseau de citations de brevets utilisé est un sous-graphe du graphe *cit-Patents* de la

Jeu de données	**# d'arêtes**	**# de noeuds**	**# d'*items* par noeud**	**# total d'*items***	**densité du graphe** [WF94]
Erosion des sols	41 166	25 618	6	262	0.000063
Réseau de citations de brevets	414 487	184 284	5-7	506	0.000012
V20K E60K λ1-5	60 000	20 000	1-5	15	0.00015
V40K E120K λ1-5	120 000	40 000	1-5	15	0.000075
V200K E600K λ1-5	600 000	200 000	1-5	15	0.000015
V20K E60K λ5-10	60 000	20 000	5-10	15	0.00015
V40K E120K λ5-10	120 000	40 000	5-10	15	0.000075
V200K E600K λ5-10	600 000	200 000	5-10	15	0.000015
V20K E60K λ5-10 en 10 couches	60 000	20 000	5-10	15	0.00015
V40K E120K λ5-10 en 10 couches	120 000	40 000	5-10	15	0.000075
V200K E600K λ5-10 en 10 couches	600 000	200 000	5-10	15	0.000015

TABLE 3.3 – Caractéristiques des différents jeux de données.

collection de jeux de données de grands réseaux de Stanford [LK14, LKF05]. Dans ce jeu de données, les noeuds représentent des brevets accordés aux États-Unis entre 1975 et 1999, et les arêtes représentent des citations. Chaque sommet est étiqueté de 5 à 7 éléments correspondant au pays, à l'état, à l'année, au type, à la catégorie et à la sous-catégorie d'un brevet.

Les neuf jeux de données synthétiques ont été générés à l'aide du programme *DigraphGenerator* proposé dans [SW11]. Cette approche construit un DAG étiqueté contenant un nombre donné de noeuds et d'arêtes. Les arêtes sont générées aléatoirement (suivant une distribution uniforme). Pour obtenir des DAGs attribués, les sommets sont simplement étiquetés avec des *itemsets*. Pour chaque noeud, le nombre d'*items* est choisi aléatoirement (suivant une distribution gaussienne). Ensuite, les *items* sont sélectionnés parmi un ensemble de quinze *items* (suivant une distribution uniforme).

3.4.3 Analyse qualitative

Plusieurs motifs intéressants pour les experts ont été extraits. La figure 3.15 présente un exemple de motif lié à l'activité minière. De nouvelles mines et des bâtiments sont apparus dans cette région entre 2005 et 2008. Nous pouvons les voir dans le centre et en bas à gauche des images 2008 et 2009. En conséquence, l'érosion du sol a augmenté dans cette région pendant cette période (Redness1 à Redness4). Cette dégradation du sol est confirmée par un faible indice de végétation (NDVI0) et un indice de luminosité élevé (Brightness4). Nous pouvons remarquer que cette dégradation s'étend progressivement. Elle a commencé en 2005 avec peu de zones (de couleur bleue dans l'image 2005) et a continué en 2008 avec d'autres à proximité (de couleur bleue dans l'image 2008). Nous observons aussi qu'une fois dégradée, la zone le reste pendant une longue période de temps (Redness4 suivi de Redness4 dans le motif).

Une petite partie de cette région a aussi été caractérisée par une augmentation de la végétation. Cette évolution a été mise en avant par le motif NDVI2 $\overset{57}{\rightarrowtail}$ NDVI3 $\overset{58}{\rightarrowtail}$ NDVI4. L'indice de végétation ($NDVI$) croît graduellement de modéré ($NDVI2$) à très élevé ($NDVI4$). Peu de zones ont suivi une telle évolution. Certaines de ces zones étaient de petits lacs ayant connu une prolifération d'algues. D'autres zones étaient de vieilles pistes abandonnées en cours de revégétalisation. Une zone est devenue une forêt. Son évolution peut être liée à une densification de la végétation due à de nouvelles plantes. Elle peut également être liée à une augmentation de la taille des arbres. Toutefois, dans un tel cas, il est étrange que d'autres forêts n'aient pas

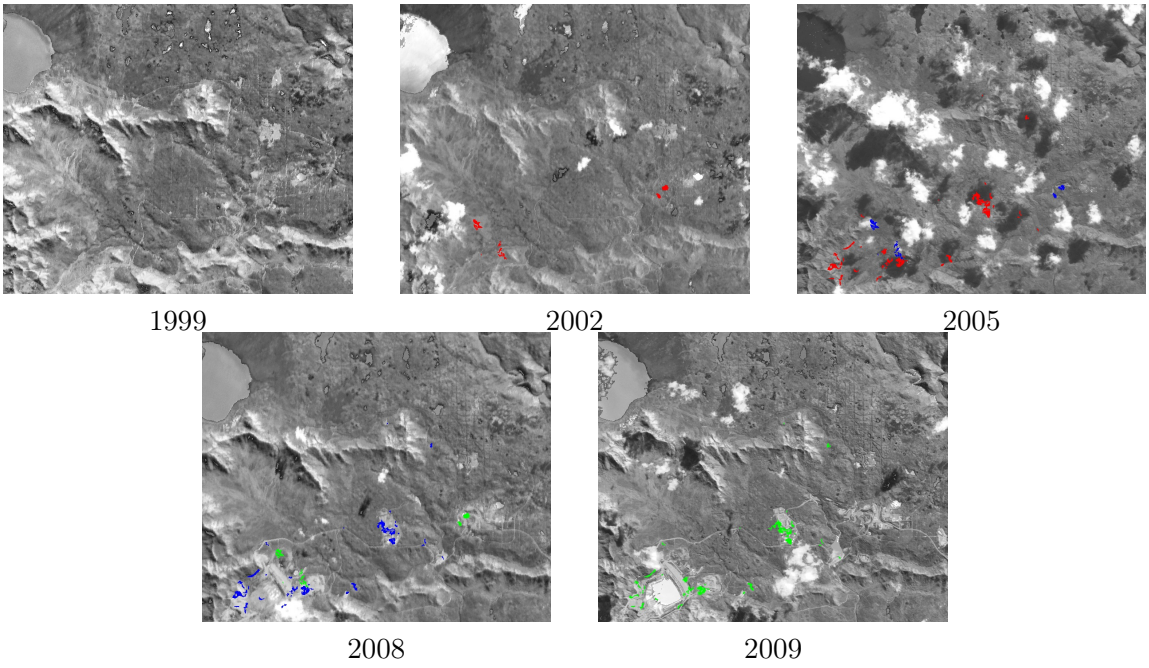


FIGURE 3.15 – Motif Redness1 $\overset{58}{\mapsto}$ Redness4, NDVI0 $\overset{61}{\mapsto}$ Redness4, NDVI0, Brightness4

connu une évolution similaire. Des investigations sur le terrain sont nécessaires pour expliquer ces changements en détail.

3.4.4 Analyse quantitative

Les expérimentations suivantes ont été effectuées sur un ordinateur avec un processeur Intel Core i5@3.10 GHz et 16 Go de mémoire principale.

Les expérimentations sur le jeu de données lié à l'érosion ont montré que l'algorithme restait efficace (moins d'une minute) jusqu'à des seuils très bas de fréquence (0.1%). Le nombre de motifs extraits quant à lui est de l'ordre de la dizaine jusqu'à un seuil de 10%. Ensuite, ce nombre croit de manière exponentielle (de l'ordre de 10000 pour un seuil de 1%). Comme le montre la figure 3.16, les performances sont similaires sur le jeu de données du réseau de citations. Sur ces données de taille beaucoup plus importante et éparses, le temps d'exécution ne dépasse pas 3 minutes jusqu'à un seuil de fréquence de 0.1% et l'espace mémoire utilisé est d'environ 1Go de mémoire. Le nombre de motifs reste relativement stable (une centaine de motifs) jusqu'à un seuil de 1%.

Les jeux de données synthétiques ont permis de confirmer ces tendances et d'étudier plus en détail les performances de notre algorithme. Par exemple, la figure 3.17 montre le temps d'exécution, l'utilisation de la mémoire et le nombre de solutions pour les trois jeux de données synthétiques ayant de 1 à 5 *items* par noeud.

La figure 3.18 illustre l'impact du nombre d'*items* par noeud sur les performances et le nombre de solutions. Comme attendu, le nombre de motifs est beaucoup plus important lorsque le nombre d'*items* augmente. Le temps d'exécution augmente donc en conséquence. Ces résultats mettent en avant la différence entre fouiller un graphe étiqueté et fouiller un graphe attribué. La complexité combinatoire est beaucoup plus élevée.

Pour finir, nous avons étudié les performances de a-DAG synthétiques ayant une structure plus proche des données spatio-temporelles. Ces a-DAGs ont le même nombre de noeuds, d'arêtes,

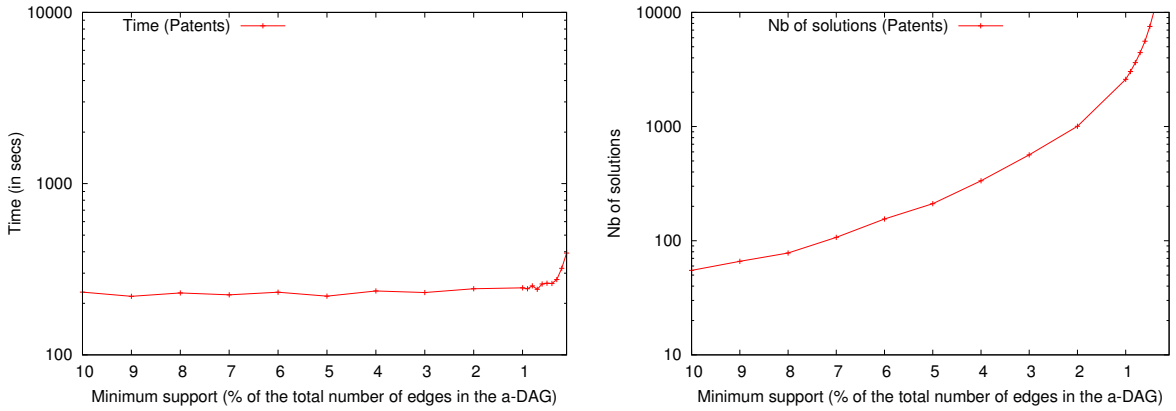


FIGURE 3.16 – Temps d'exécution et nombre de solutions pour le réseau de citations.

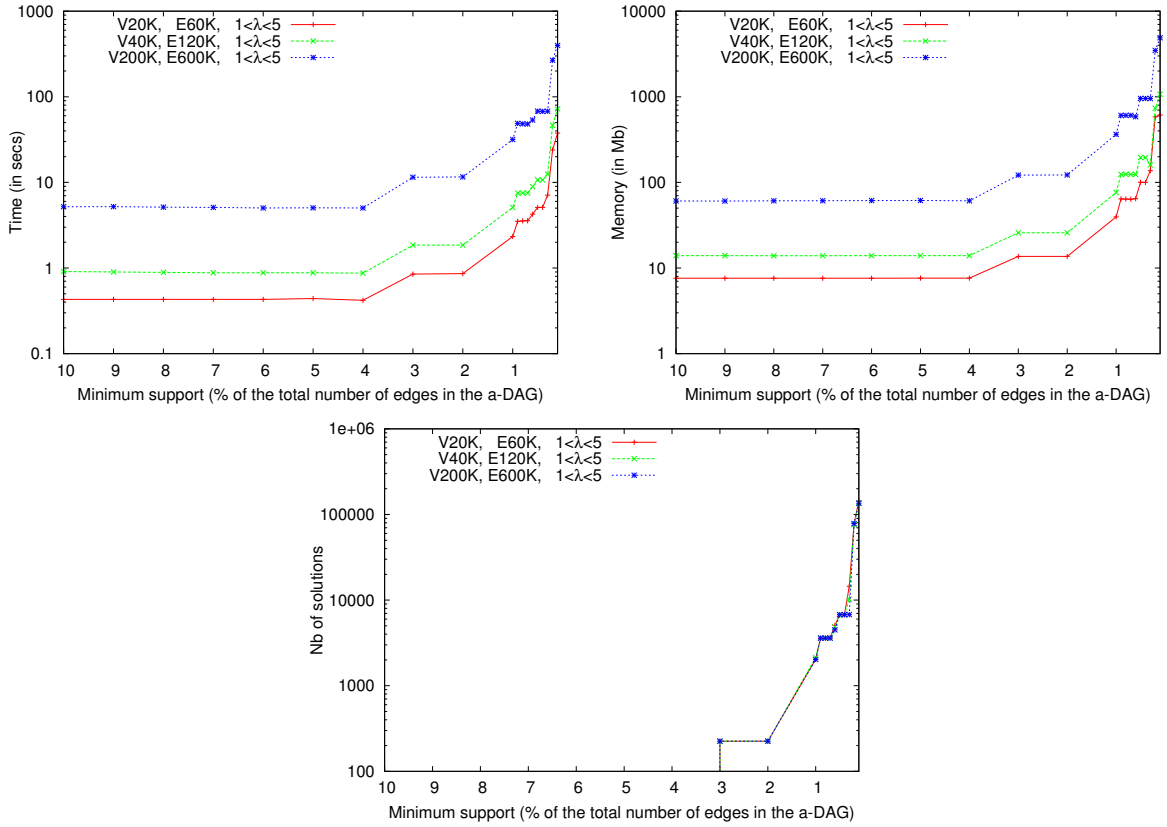


FIGURE 3.17 – Temps d'exécution, utilisation mémoire, et nombre de solutions pour des a-DAG synthétiques de différentes tailles.

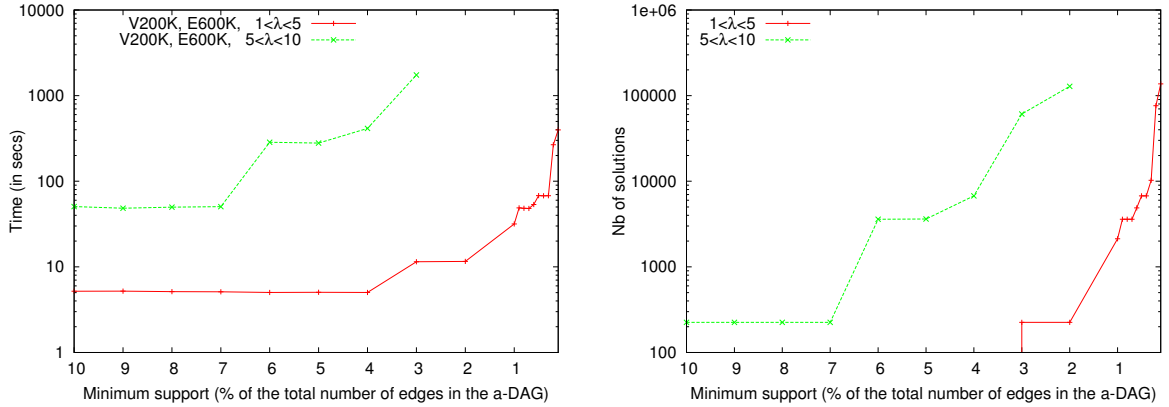


FIGURE 3.18 – Temps d'exécution et nombre de solutions pour des a-DAG synthétiques lorsque le nombre d'*items* par noeud augmente.

et d'*itemsets*. Toutefois, les noeuds sont répartis en dix ensembles (appelés "couches") organisés en série temporelle. De plus, les arêtes sont uniquement entre des noeuds de couches consécutives. Ces graphes ont donc des chemins plus longs et un plus grand nombre de motifs sont extraits. Comme illustré par la figure 3.19, les temps d'exécution et l'utilisation de la mémoire sont beaucoup plus élevés avec ce type de structure, mais reste acceptable jusqu'à un seuil de fréquence 7 %.

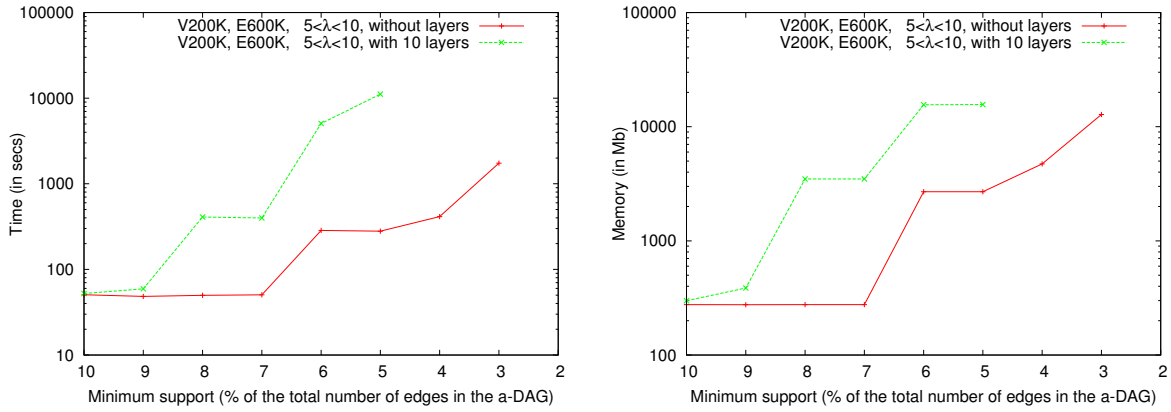


FIGURE 3.19 – Temps d'exécution et utilisation mémoire des a-DAG synthétiques avec une structure en couches.

4

Extraction de motifs récurrents dans des graphes dynamiques attribués

L'extraction de chemins pondérés dans un unique *DAG* attribué permet de mettre en avant des évolutions fréquentes d'objets spatiaux avec des dynamiques complexes. Toutefois, cette représentation sous forme de *DAG*, et ce domaine de motifs, ne considèrent que très partiellement la dimension spatiale. Ils ne permettent pas par exemple de prendre en compte le voisinage d'une évolution, qui pourtant pourrait l'expliquer.

Face à cette limite, une représentation des données plus complète d'un point de vue spatio-temporel doit être considérée. Dans le domaine du traitement d'images, les différents objets d'une image et leurs relations spatiales sont communément représentés sous forme de graphe. Nous avons donc adopté cette structure pour représenter la dimension spatiale des données. Nous considérons plus précisément des graphes attribués pour capturer toutes les informations associées aux objets spatiaux. Au final, l'évolution temporelle de ces objets spatiaux est donc modélisée par une séquence de graphes attribués, encore appelée **graphe dynamique attribué**. Comme discuté en section 2.1.2, l'extraction de motifs dans de telles structures de données est encore peu étudié dans la littérature. Les travaux existants se focalisent sur des sous-graphes dont les noeuds évoluent de manière identique [DPRB12, DPRB13] ou sur des évolutions topologiques [KPPR15]. Dans ce chapitre, nous proposons donc un domaine de motifs plus générique permettant d'extraire des évolutions récurrentes de sous-graphes connexes. Ces motifs représentent des séquences d'ensembles de caractéristiques liés spatialement et apparaissant plusieurs fois au même endroit à des dates différentes.

Ce travail a été réalisé dans le cadre d'une thèse du Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. Il est centré sur des aspects méthodologiques et n'est pas associé à un domaine d'application particulier. Il a été appliqué à différents jeux de données réels disponibles dans la communauté (impact de cyclones sur le trafic aérien aux Etats-Unis et analyse du réseau de publications DBLP) et à des jeux de données synthétiques. Dans le cadre de sa thèse, le doctorant a aussi appliqué ce travail au suivi de bassins aquacoles à partir d'une série d'images satellitaires (en collaboration avec l'Ifremer).

Le tableau 1.1 présente le doctorant ayant travaillé sur cette problématique, ainsi que les projets de recherche et les collaborations associés. Suite à ce tableau sont également listées les principales publications.

Ce chapitre est organisé de la manière suivante. La section 3.1 présente le domaine de motifs proposé et les contraintes utilisées. La section 3.2 décrit une stratégie originale pour extraire ces motifs dans un unique graphe dynamique attribué. Pour finir, la section 3.4 discute des résultats

obtenus sur différents jeux de données synthétiques et réels.

Master/Thèse	**Collaborations**
Z. Cheng (2014-2018) thèse dirigée par N. Selmaoui-Folcher (UNC), co-encadrant : F. Flouvat	Ifremer

TABLE 4.1 – Synthèse des encadrements et des collaborations en lien avec l'extraction de motifs récurrents dans un graphe dynamique attribué

Principales publications
Zhi Cheng, Frédéric Flouvat and Nazha Selmaoui-Folcher. Mining reccurrent patterns in a dynamic attributed graph. In Proceedings of the 21st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PaKDD'17), Jeju, South Korea, 2017.

TABLE 4.2 – Synthèse des publications en lien avec l'extraction de motifs récurrents dans un graphe dynamique attribué

4.1 Cadre théorique

4.1.1 Les données

Dans ce chapitre, la base de données en entrée est un **unique graphe dynamique attribué** $\mathcal{G} =< G_{t_1}, G_{t_2}, ..., G_{t_{max}} >$ représentant l'évolution d'un graphe sur un ensemble de temps $\mathcal{T} = \{t_1, t_2, ...t_{max}\}$. L'ensemble des noeuds de $\mathcal{G}$ est noté $\mathcal{V}$. Un noeud de $\mathcal{G}$ est étiqueté par des *items* de l'ensemble $\mathcal{I}$. A chaque temps $t \in T$, le graphe $\mathcal{G}$ est un graphe attribué non-orienté noté $G_t = (V_t, E_t, \lambda_t)$ où $V_t \subseteq \mathcal{V}$ est l'ensemble des noeuds au temps t , $E_t \subseteq V_t \times V_t$ est l'ensemble des arêtes au temps t, et $\lambda_t : V_t \rightarrow \mathcal{P}(\mathcal{I})$ est la fonction associant chaque sommet de V_t à un sous-ensemble d'*items* de $\mathcal{I}$. La figure 4.1 présente un exemple de graphe dynamique attribué.

Ce type de représentation peut être utilisée pour modéliser des données variées. Par exemple, l'évolution d'un réseau social peut être représentée par un tel graphe, tout comme l'évolution d'un ensemble d'objets spatiaux. Dans le premier cas, chaque noeud correspond à une personne, son étiquette est un ensemble d'informations (*items*) sur la personne sous-jacente, et les arêtes correspondent à la relation sociale entre les personnes. Dans le second cas, chaque noeud représente un objet spatial, son étiquette est un ensemble d'informations sur l'objet associé, et les arêtes représentent les relations spatiales (p.ex. le voisinage). Dans tous les cas, les noeuds, les étiquettes et les arêtes peuvent varier au cours de temps (modification, apparition ou disparition).

4.1.2 Les évolutions récurrentes

Dans ce travail, nous avons introduit un nouveau domaine de motifs, appelé motifs récurrents, pour analyser ces graphes. Ils correspondent à des évolutions récurrentes des informations associées aux étiquettes des noeuds, i.e. des sous-séquences de sous-graphes attribués. Les arêtes des sous-graphes ne sont pas prises en compte directement afin d'élargir l'analyse. Les motifs sont composés de sous-graphes "sans arêtes". Par contre, elles sont prises en compte dans plusieurs contraintes structurelles sur les sous-graphes de la séquence (p.ex. connectivité). Ces contraintes seront détaillées dans la section suivante.

Plus formellement, soit (V, λ) un sous-ensemble de noeuds attribués de $\mathcal{G}$ avec $V \subseteq \mathcal{V}$ et $\lambda : V \rightarrow \mathcal{P}(\mathcal{I})$. (V, λ) peut être vu comme un graphe attribué sans arêtes. Pour faciliter la

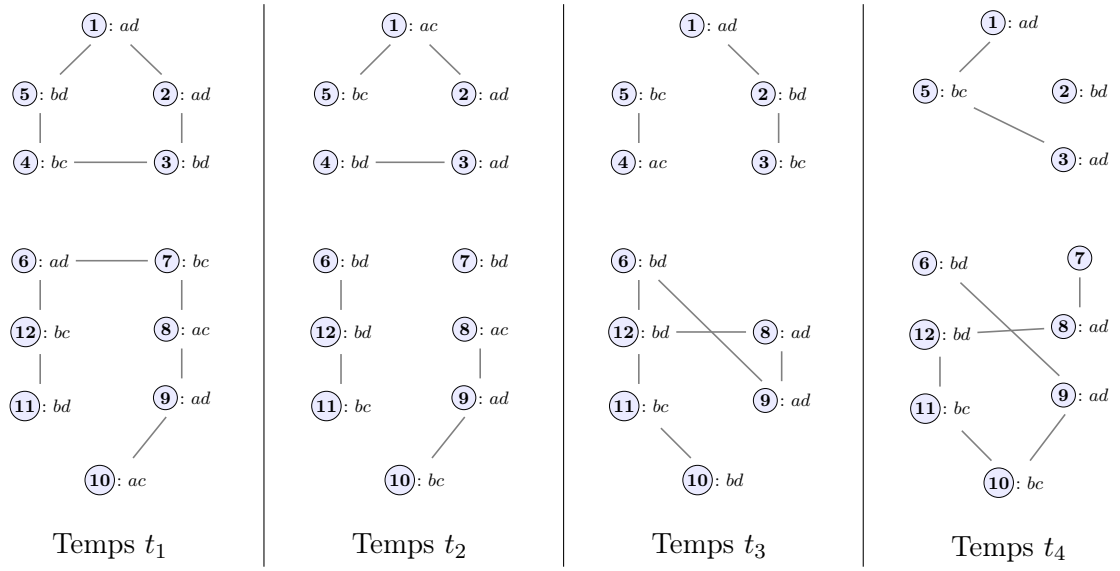


FIGURE 4.1 – Exemple de graphe dynamique attribué

lecture des exemples dans le texte, l'ensemble de noeuds attribués (V, λ) pourra aussi être noté $(v_1 : \lambda(v_1) \mid v_2 : \lambda(v_2) \mid ...)$, $\forall v_1, v_2... \in V$. Comme montre la figure 4.1, $(1 : ad \mid 2 : ad \mid 3 : bd \mid 4 : bc \mid 5 : bd)$ est un ensemble de noeuds attribués de t_1.

L'évolution d'un sous-ensemble de noeuds de $\mathcal{G}$ à un temps $t \in \mathcal{T}$ est une séquence $S =< (V'_1, \lambda'_1), (V'_2, \lambda'_2), ..., (V'_k, \lambda'_k) >$ tel que $\forall i \in \{1, 2, ..., k\}$, $\exists E'_i \subseteq E_{t+i-1}$ et (V'_i, E'_i, λ'_i) est un sous-graphe de G_{t+i-1}. Par exemple, dans la figure 4.1, $\langle (1 : a \mid 2 : d \mid 3 : bd \mid 4 : bc \mid 5 : bd)$ $(1 : ac \mid 2 : ad \mid 5 : bc)\rangle$ est une évolution débutant au temps t_1.

Soit $T_P = \{t_{i_1}, t_{i_2}, ..., t_{i_m}\}$ un ensemble de temps associés à l'évolution $S_P =< (V'_1, \lambda'_1), (V'_2, \lambda'_2), ..., (V'_k, \lambda'_k) >$. Une **évolution récurrente d'un sous-ensemble de noeuds** de $\mathcal{G}$ à l'ensemble de temps T_P, selon la séquence S_P, est notée $P = (S_P, T_P)$. Dans ce cas, la taille de P est k. Dans la figure 4.1, $\big(\langle (1 : a \mid 2 : ad \mid 5 : b)(1 : a \mid 2 : d)\rangle,\ \{t_1, t_2\}\big)$ est un exemple d'évolution récurrente débutant aux temps t_1 et t_2. Ce motif représente une sous-graphe connexe composé de v_1, v_2, et v_5 avec les attributs $\{a, ad, b\}$, suivi le temps suivant par un sous-graphe composé de v_1 et v_2 avec les attributs $\{a, d\}$.

4.1.3 Les contraintes : structure, temporalité, redondance et fréquence

Plusieurs contraintes sont utilisées pour définir si un motif $P = \big(\langle (V'_1, \lambda'_1) \dots (V'_k, \lambda'_k)\rangle,\ T_P\big)$ est intéressant. Nous avons tout d'abord considéré la structure des sous-graphes par l'intermédiaire de contraintes de connectivité, de cohésion et de volume. Ensuite, nous avons intégré une contrainte temporelle de continuité. Pour finir, les contraintes de fréquence minimale et de non-redondance ont aussi été adaptées à ce nouveau contexte.

Connectivité. Les noeuds d'un graphe représentent souvent des individus/objets, et les arêtes représentent des relations entre ces individus/objets (p.ex. relation spatiale, relation d'amitié). L'intégration d'une contrainte de connectivité (ou connexité) entre les noeuds permet donc de se focaliser sur des évolutions potentiellement corrélées. Par exemple, dans la figure 4.1, $\big(\langle (1 : a \mid 2 : ad \mid 5 : b)(1 : a \mid 2 : d)\rangle,\ \{t_1, t_2\}\big)$ représente une évolution de noeuds connexes, et donc

potentiellement liés. Cette contrainte peut être définie de la manière suivante :

$\mathbf{Q_{conn}} \equiv \forall t \in T_P, \forall i \in \{1, 2, ..., k\}$, les noeuds de V'_i forment une composante connexe dans G_{t+i-1}

Cohésion. Cette contrainte utilisée dans [DPRB12] garantit la cohésion des noeuds du motif à un temps donné. Elle s'appuie sur la similarité du voisinage pour extraire des noeuds qui sont étroitement liés. Par exemple, cette contrainte permet de considérer des objets qui sont spatialement très proches. Etant donné un seuil de similarité $minsim \in [0, 1]$ et une mesure de similarité $sim()$, la contrainte de cohésion minimale de P est

$$\mathbf{Q_{cohe}} \equiv \forall v \in V'_i, 1 \leq i \leq k, \exists u \in V'_i, \text{ tel que } sim(v, u, V'_i) \geq minsim$$

Toute mesure de similarité peut être utilisée pour centrer l'extraction sur des motifs représentant différentes structures de graphe. Il est notamment possible d'utiliser les mesures de similarité *Cosine* [T+06] et *Jaccard* [Jac12] qui considèrent la similarité du voisinage direct des noeuds.

Volume. Le volume est une autre mesure couramment étudiée lors de l'analyse de graphes. Elle correspond au nombre minimum de noeuds des sous-graphes considérés. Par exemple, le motif $\big(\langle(1 : a \mid 2 : ad \mid 5 : b)(1 : a \mid 2 : d)\rangle, \{t_1, t_2\}\big)$ a un volume de 2. Cette mesure représente la taille d'une communauté dans un réseau social (en supposant que les noeuds soient les individus et les arêtes les liens d'amitié). Etant donné un seuil de volume *minvol*, la contrainte de volume minimum de P est

$$\mathbf{Q_{vol}} \equiv min_{\forall i \in \{1...k\}}(|V'_i|) \geq minvol$$

Continuité temporelle. Par défaut, une évolution peut représenter des noeuds totalement différents à chaque temps. L'interprétation par les utilisateurs de telles évolutions peut être difficile car il n'y a pas a priori de liens directs (pas d'arêtes) entre les individus/objets observés (représentés par les noeuds). Nous proposons donc une nouvelle contrainte visant à cibler les motifs décrivant des évolutions autour d'un noyau commun d'individus. Par exemple, le motif $\big(\langle(1 : a \mid 2 : ad \mid 5 : b)(1 : a \mid 2 : d)\rangle, \{t_1, t_2\}\big)$ a deux noeuds communs aux temps t_1 et t_2. Ainsi, il est possible de suivre l'évolution dans le temps d'un certain nombre de noeuds, tout en considérant aussi les noeuds voisins (directement ou indirectement). Etant donné un nombre de noeuds seuil *mincom* défini par l'utilisateur, cette contrainte peut être formalisée de la manière suivante :

$$\mathbf{Q_{cont}} \equiv |\bigcap_{\forall i \in 1...k} V'_i| \geq mincom$$

Non-redondance. Tout comme dans le travail précédent sur les a-DAG, nous avons intégré une contrainte visant à éliminer les motifs redondants, i.e. portant la même information. Par exemple, $P = \big(\langle(1 : a \mid 2 : ad)(1 : a \mid 2 : d)\rangle, \{t_1, t_2\}\big)$ est redondant par rapport à $P2 = \big(\langle(1 : a \mid 2 : ad \mid 5 : b)(1 : a \mid 2 : d)\rangle, \{t_1, t_2\}\big)$. En effet, la séquence de P est incluse dans celle de $P2$ et les deux motifs apparaissent exactement aux mêmes moments. Soit $Th(\mathcal{G}, Q)$ l'ensemble des motifs solutions dans $\mathcal{G}$ par rapport à une contrainte Q. La contrainte de non-redondance de P est

$$\mathbf{Q_{nonRedund}} \equiv P \in Th(\mathcal{G}, Q) \mid \nexists P2 \in Th(\mathcal{G}, Q) \text{ tel que } P \sqsubseteq P2 \text{ et } T_P = T_{P2}$$

Fréquence. Classiquement, la fréquence représente soit le nombre de "transactions" contenant le motif, soit son nombre d'occurrences. Dans notre cas, la définition de fréquence est différente. Elle représente le nombre d'apparition du motif dans le temps, i.e. le nombre de temps auquel l'évolution commence. Par exemple, dans la figure 4.1, la fréquence de $\big(\langle(6:d \mid 11:b \mid 12:b)(11:bc \mid 12:bd)\rangle, \{t_1, t_2, t_3\}\big)$ est 3 puisque cette évolution débute à t_1, t_2 et t_3. Plus formellement, le motif P est fréquent dans $\mathcal{G}$ par rapport à un seuil minimum *minsup* si

$$\mathbf{Q_{freq}} \equiv |T_P| \geq minsup$$

4.1.4 Problématique

Etant donné un graphe dynamique attribué $\mathcal{G}$, l'objectif est d'énumérer l'ensemble des évolutions récurrentes dans $\mathcal{G}$ vérifiant la contrainte $Q = Q_{conn} \wedge Q_{cohe} \wedge Q_{vol} \wedge Q_{cont} \wedge Q_{nonRedund} \wedge Q_{freq}$, noté $Th(\mathcal{G}, Q)$.

4.2 Stratégie d'extraction des motifs récurrents

Nous avons introduit une stratégie originale pour extraire ces motifs. Elle ne s'appuie pas sur une stratégie générer-tester (où des motifs candidats sont générés, testés, puis combinés). Elle ne suit pas un parcours en largeur ou en profondeur de l'espace de recherche. Elle ne fait pas non plus un parcours basé sur des projections successives des données (tel que le font par exemple *PrefixSpan* et ses autres variantes). Elle s'appuie sur des intersections successives des composantes connexes contenues à chaque temps. Au fur et à mesure de ces intersections et du parcours des temps, les motifs sont progressivement étendus. On obtient ainsi à chaque itération (à chaque temps) un ensemble de motifs solutions de taille potentiellement différente. L'avantage de cette approche est d'éviter la génération d'un grand nombre de motifs ne vérifiant pas les contraintes et de traiter de manière incrémentale le graphe dynamique attribué, i.e. la séquence de graphes, en entrée. La sous-section suivante va introduire la notion d'intersections entre des graphes attribués. Cette notion est à la base de la stratégie d'énumération proposée.

4.2.1 Intersections de graphes attribués et motifs de taille 1

Lien entre intersection et fréquence Considérons deux temps $i, j \in \mathcal{T}$. L'intersection entre les deux graphes attribués $G_i = (V_i, E_i, \lambda_i)$ et $G_j = (V_j, E_j, \lambda_j)$, notée $G_i \sqcap G_j$, est un graphe attribué $G = (V, E, \lambda)$ tel que $V = V_i \cap V_j$, $E = E_i \cap E_j$, $\forall v \in V$, $\lambda(v) = \lambda_i(v) \cap \lambda_j(v)$. Autrement dit, il s'agit d'un graphe constitué des noeuds, des arêtes et des valeurs d'attributs communs aux deux graphes en entrée. On remarque que tout sous-graphe de G apparaît au temps i et j. Ils apparaissent donc au moins deux fois dans $\mathcal{G}$. La figure 4.2 présente un exemple illustrant l'intersection de trois graphes $(G_1 \sqcap G_3 \sqcap G_4)$. Le sous-graphe c appartenant à cette intersection a la propriété d'apparaître au moins 3 fois dans les données (au moins dans t_1, t_3 et t_4).

Cette propriété peut être généralisée à l'intersection de k graphes, avec $k \in \{2, 3, ...|\mathcal{T}|\}$. Soit $T^k \subseteq \mathcal{T}$ un sous-ensemble de temps de $\mathcal{G}$ tel que $|T^k| = k$. L'intersection des graphes de $\mathcal{G}$ sur les k temps de T^k, noté $\underset{i \in T^k}{\sqcap} G_i$, est un graphe $G = (V, E, \lambda)$ avec $V = \underset{i \in T^k}{\cap} V_i$, $E = \underset{i \in T^k}{\cap} E_i$, $\forall v \in V$, $\lambda(v) = \underset{i \in T^k}{\cap} \lambda_i(v)$. La fréquence minimale d'apparition dans $\mathcal{G}$ de tout sous-ensemble de noeuds attribués de $\underset{i \in T^k}{\sqcap} G_i$ est k. Ainsi, tout motif construit à partir de l'intersection de *minusp* graphes de $\mathcal{G}$ respectera la contrainte de fréquence minimale.

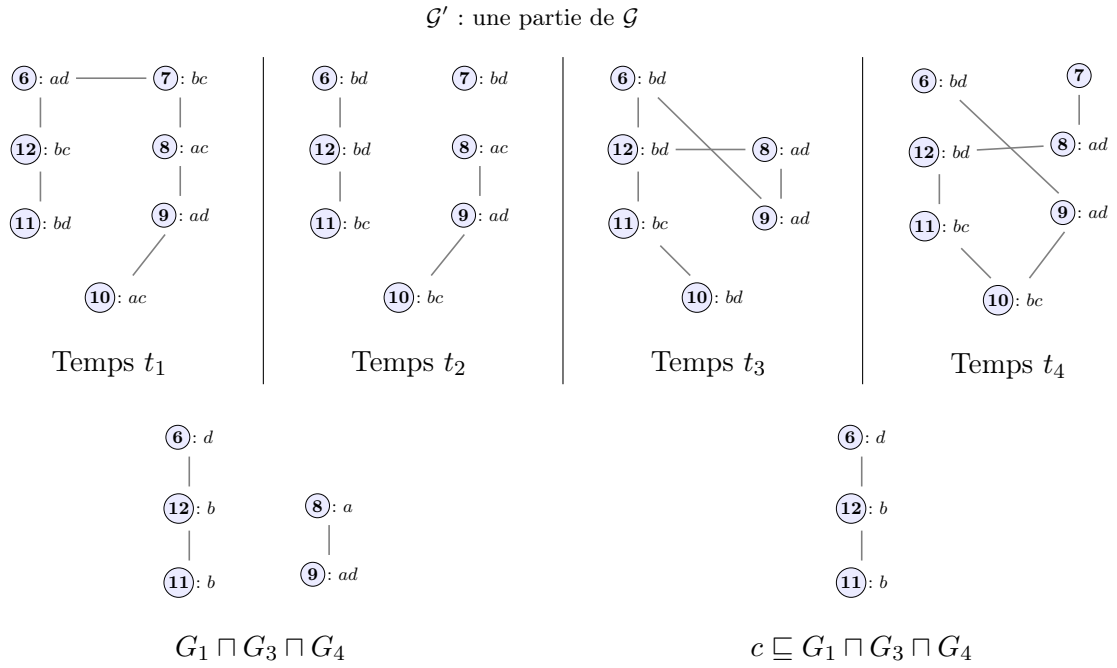


FIGURE 4.2 – Exemple d'intersection de graphes

Lien entre intersection et non-redondance L'intersection de graphes permet aussi d'avoir d'autres propriétés. Etudions plus particulièrement les composantes connexes (i.e. les sous-graphes connectés maximaux) issues de l'intersection de plusieurs graphes. Nous noterons $\mathbb{C}_{i \sqcap j}$ l'ensemble des composantes connexes du graphe obtenu après intersection du graphe $\mathcal{G}$ aux temps i et j, i.e. $G_i \sqcap G_j$.

Considérons deux composantes connexes c et c' obtenues après intersection des graphes aux temps $\{i, j\}$ et $\{k, l\}$, i.e. $c \in \mathbb{C}_{i \sqcap j}$ et $c' \in \mathbb{C}_{k \sqcap l}$, $\forall i, j, k, l \in \mathcal{T}$. Soit $T_c = \{t \in \mathcal{T} \mid c \sqsubseteq G_t\}$ (resp. $T_{c'}$) l'ensemble des temps de $\mathcal{T}$ où la composante connexe c (resp. c') apparaît. Il est impossible d'avoir $c \sqsubset c'$ (ou l'inverse) et $\boldsymbol{T_c = T_{c'}}$. Dans la figure 4.2, la composante connexe $(6 : d \mid 12 : b \mid 11 : b)$ est dans G_1, G_3 et G_4. Elle apparaît donc dans $G_1 \sqcap G_3$ et $G_1 \sqcap G_4$. Il n'existe pas de sur-ensemble de noeuds attribués l'incluant et apparaissant aux mêmes temps. A noter que le sous-ensemble $(12 : b \mid 11 : b)$ est obtenu en faisant $G_1 \sqcap G_2$, mais il apparaît à des temps différents (t_1, t_2, t_3 et t_4). Pour résumer, si $c = (V, E, \lambda)$, alors le motif $(< (V, \lambda) >, T_c)$ vérifie la contrainte de connexité (car c est une composante connexe), mais aussi la contrainte de non-redondance (dans l'ensemble des solutions de taille 1). Autrement dit, ce motif sera donc soit un motif solution, soit une partie d'un motif solution.

Cette propriété peut être généralisée à tout ensemble $T, T' \subseteq \mathcal{T}$. Notons $\mathbb{C}_{\sqcap T}$ (resp. $\mathbb{C}_{\sqcap T'}$), l'ensemble des composantes connexes obtenues après intersection des graphes aux temps T (resp. T'), i.e $\underset{i \in T}{\sqcap} G_i$. Si $c \in \mathbb{C}_{\sqcap T}$, alors $\nexists c' \in \mathbb{C}_{\sqcap \mathbb{T}'}$ tel que $c \sqsubset c'$ et $T_c = T_{c'}$. Les motifs de taille 1 associés à ces composantes connexes vérifient donc les contraintes de connectivité et de non-redondance. Pour que ces motifs soient des solutions, il suffit de vérifier en plus les contraintes de volume et de continuité (des contraintes peu coûteuses d'après leur nature). La réciproque est aussi vraie. Tout motif solution de taille 1, ou tout sous-motif de taille 1 d'un motif solution (i.e. une "partie" d'un motif solution), peut être dérivé des composantes connexes obtenues après intersections entre des graphes de $\mathcal{G}$. Ces intersections nous permettent donc d'obtenir les "briques de bases" servant à construire l'ensemble des motifs solutions.

L'intérêt de ces intersections est d'éviter d'avoir à faire un grand nombre de tests d'inclusion lors de l'extraction (pour vérifier les contraintes de fréquence et de non-redondance), et donc de gagner en efficacité. En effet, le nombre d'intersections est $2^{|\mathcal{T}|}$ alors que le nombre de tests d'inclusion dépend du nombre de motifs générés. Ce dernier est donc dans le pire cas, et en général, beaucoup plus important (de plusieurs ordres de grandeur). De plus, même si le nombre d'opérations pour faire une intersection est en moyenne plus important que pour faire un test d'inclusion, la complexité des algorithmes reste du même ordre de grandeur.

Par ailleurs, un prétraitement est aussi effectué afin de faire ces intersections plus rapidement. Il consiste à rechercher toutes les composantes connexes de G_i $(1 \leq i \leq |\mathcal{T}|)$ dont les volumes sont supérieurs à *minvol*, notés $\mathbb{C}_i$. Dans la figure 4.1, l'algorithme recherche donc d'abord l'ensemble des composantes connexes de G_1, i.e., $\mathbb{C}_1 = \{(v_1, v_2, v_3, v_4, v_5), (v_6, v_7, v_8, v_9, v_{10}, v_{11}, v_{12})\}$. Puis, $\mathbb{C}_2 = \{(v_1, v_2, v_5), (v_3, v_4), (v_6, v_{11}, v_{12}), (v_8, v_9, v_{10})\}$, $\mathbb{C}_3 = \{(v_1, v_2, v_3), (v_4, v_5), (v_6, v_7, v_8, v_9, v_{10}, v_{11}, v_{12})\}$ et $\mathbb{C}_4 = \{(v_1, v_3, v_5), (v_6, v_7, v_8, v_9, v_{10}, v_{11}, v_{12})\}$ sont identifiés. Ensuite, les intersections ne sont plus faites sur le graphe initial mais entre leurs composantes connexes, ce qui évite des tests de connexité et permet d'éliminer directement les composantes ne vérifiant pas la contrainte de volume minimum.

4.2.2 Génération incrémentale des motifs

Les motifs de taille 1 extraits dans les intersections peuvent ensuite être combinés, en fonction des temps où ils apparaissent, afin de générer les autres motifs. Cette extension se fait de manière incrémentale en traitant les temps les uns après les autres.

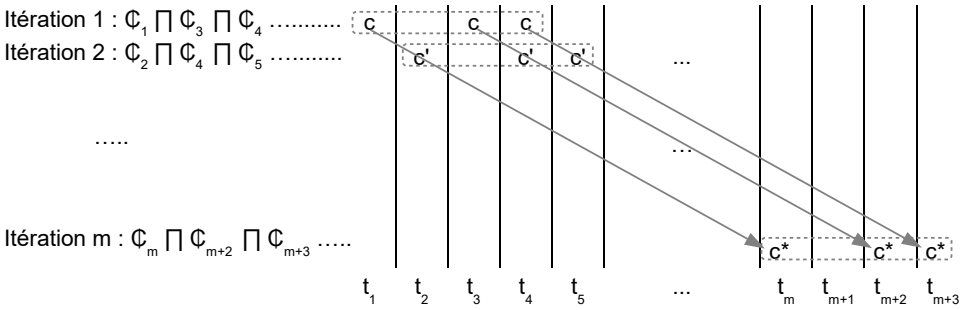


FIGURE 4.3 – Extension en parallèle des motifs de type ($< c, c', ..., c^* >$, $\{t_1, t_3, t_4\}$) à partir de $\{t_1, t_3, t_4\}$

La figure 4.3 illustre cette construction incrémentale à partir des temps t_1, t_3 et t_4. Elle représente l'extension en parallèle des motifs apparaissant en t_1, t_3 et t_4 (de trois de leurs occurrences donc). A noter que la contrainte de fréquence étant directement liée au nombre de temps "intersectés", on peut déduire que la fréquence minimale dans cet exemple est 3. Supposons par exemple qu'il existe un motif solution $P = (< (V'_1, \lambda'_1), (V'_2, \lambda'_2), ...(V'_n, \lambda'_n) >, \{t_1, t_3, t_4\})$. Les traitements réalisés pour $\{t_1, t_3, t_4\}$ permettent d'obtenir trois occurrences de la séquence $< (V'_1, \lambda'_1) >$ (celles en t_1, t_3 et t_4). L'occurrence en t_1 ne peut être étendue que par un ensemble de noeuds attribués de t_2 (aucun "*gap*" de temps n'étant autorisé), à condition que ces deux ensembles aient un nombre suffisamment important de noeuds en commun (contrainte de continuité). De même, celle en t_3 ne peut être étendue que par les motifs de taille 1 trouvés en t_4 (de même pour t_4 et t_5). A partir de l'intersection de $\{t_2, t_4, t_5\}$, on peut étendre $< (V'_1, \lambda'_1) >$ avec (V'_2, λ'_2) issu de cette intersection, et obtenir $< (V'_1, \lambda'_1), (V'_2, \lambda'_2) >$. Ce processus continue jusqu'à ce que l'on ne puisse plus étendre les séquences étudiées. On obtient alors un motif so-

lution (non-redondant). Au delà du motif P, cette succession d'extensions permet de générer incrémentalement tous les motifs commençant en t_1, tous les motifs commençant en t_2, etc. En effet, si un motif de taille 1 obtenu à partir de $\{t_2, t_4, t_5\}$ ne peut être utilisé pour étendre un motif construit "au temps précédent", il constitue le point de départ pour un nouveau motif.

Avec cette approche, le motif P peut être généré et étendu potentiellement 4 fois (à partir de $\{t_1, t_3\}$, de $\{t_1, t_4\}$, de $\{t_3, t_4\}$, et de $\{t_1, t_3, t_4\}$). A chaque génération, les temps associés au motif sont mis à jour. Bien que l'étude de la combinaison $\{t_1, t_2, t_3\}$ n'apporte pas d'informations par rapport P, elle peut permettre de découvrir ou d'étendre d'autres motifs. Toutes ces combinaisons d'intersections sont donc nécessaires, d'où l'importance du pré-traitement décrit précédemment pour limiter le coût de cette opération.

L'algorithme 4 présente en détail l'approche développée. La ligne 1 correspond à l'extraction des composantes connexes de chaque graphe. Les lignes 3-8 construisent les motifs de taille 1 dont la fréquence est supérieure au seuil minimum et dont au moins une occurrence commence au temps t_1. Pour cela, l'algorithme calcule toutes les combinaisons de temps contenant t_1 (T_1^k, ligne 4), puis génère des motifs de taille 1 en faisant l'intersection des composantes connexes apparaissant à ces temps (ligne 6, méthode *ExtractIntersect*). Les autres temps sont ensuite traités les uns après les autres. Pour chaque temps t_i, on construit de nouveau toutes les combinaisons de temps (non déjà traitées) contenant t_i (T_i^k, ligne 12), et on extrait des motifs P_i de taille 1 de l'intersection des composantes connexes (ligne 13). On essaye ensuite d'étendre chaque motif P généré à l'itération précédente (ligne 14-15). Si le motif P' résultant de l'extension de P avec P_i vérifie la contrainte de continuité, on l'ajoute dans les motifs générés au temps t_i (ligne 16-17). Sinon, on enregistre dans les solutions le motif P généré à l'étape précédente et on enregistre P_i pour une future extension. A la fin (ligne 27), l'algorithme réunit les motifs solutions construits à chaque temps.

4.3 Expérimentations et applications

4.3.1 Prototype

Ce travail n'a pas encore fait l'objet d'un développement poussé d'un point de vue interface utilisateur ni d'une intégration dans une plate-forme d'analyse de données telle que KNIME. L'algorithme a été implémenté en C++ à partir de librairies de traitements de graphes telles que *Lemon* [DJK11] et *Boost Graph* [SLL01]. Le prétraitement des données et la visualisation des motifs ont quant à eux été réalisés à partir de scripts MATLAB.

4.3.2 Protocole expérimental

Le domaine de motifs et l'approche proposée ont été appliqués à trois jeux de données réels : $DBLP$, trafic aérien aux USA et aquaculture en Indonésie. Les deux premiers jeux de données sont issus de travaux de la littérature liés à l'analyse d'un graphe dynamique attribué [DPRB12, KPPR15]. L'objectif de ces jeux de données est de comparer indirectement les différentes approches (même si les domaines de motifs sont différents). Le premier jeu de données correspond à une partie du réseau de citations $DBLP$. Ce graphe représente des auteurs et leurs co-publications entre 1990 et 2009 (subdivisées en 9 périodes au total). Les 2 723 noeuds représentent des auteurs ayant plus de 10 publications, et ils sont étiquetés par 43 attributs représentant leurs nombres de publications dans 43 conférences du domaine. Il y a 10 737 arêtes en moyenne à chaque temps et elles représentent les auteurs ayant publié ensemble. Le deuxième jeu de données représente le trafic aérien aux Etats-Unis après l'ouragan Katrina (8 dates/semaine).

Algorithm 4 *RPMiner* : extraction des évolutions récurrentes

Require: un graphe dynamique attribué $\mathcal{G}$, $minsup$: seuil minimum de fréquence, $minvol$: seuil minimum de volume, $mincom$: nombre minimum de noeuds communs au cours du temps, $minsim$: seuil minimum de cohésion
Ensure: $Th(\mathcal{G}, Q)$: l'ensemble des évolutions récurrentes vérifiant la contrainte $Q = Q_{conn} \wedge Q_{cohe} \wedge Q_{vol} \wedge Q_{cont} \wedge Q_{nonRedund} \wedge Q_{freq}$

$\mathbb{C} = \{\mathbb{C}_i$ ensemble des composantes connexes de $G_i \mid \forall c \in \mathbb{C}_i, c = (V, E, \lambda), |V| \geq minvol, \forall v \in V, \exists u \in V$ tel que $cosine(v, u) \geq mincos\}$
$Cand_i = \emptyset, \forall i \in \{1, 2, ..., |\mathcal{T}|\}$
for $k = minsup$ to $|\mathcal{T}|$ **do**
  $T_i^k = \{t_{j_1}, ..., t_{j_k} | t_{j_1} < t_{j_k}$ et $t_{j_1} = t_i\}$
  for chaque $T \subseteq T_1^k$ **do**
    $Cand_1 = Cand_1 \cup \{P_1 \in ExtractIntersect(\mathbb{C}, T)\}$
  end for
end for
$Sol_i = \emptyset, \forall i \in \{1, 2, ..., |\mathcal{T}|\}$
for $i = 2$ to $|\mathcal{T}|$ **do**
  for $k = minsup$ to $|\mathcal{T}|$ **do**
    for chaque $T \subseteq T_i^k$ **do**
      for chaque $P_i \in ExtractIntersect(\mathbb{C}, T)$ **do**
        for chaque $P = (S, T_P)$ tel que $P \in Cand_{i-1}$ and $T_P = T$ **do**
          $P' = ExtendWith(P, P_i)$
          if $Q_{cont}(P')$ **then**
            $Cand_i = Cand_i \cup \{P'\}$
          else
            $Sol_{i-1} = Sol_{i-1} \cup \{P\}$
            $Cand_i = Cand_i \cup \{P_i\}$
          end if
        end for
      end for
    end for
  end for
end for
$Th(\mathcal{G}, Q) = MergeUpdate(\bigcup_{\forall i \in \mathcal{T}} Sol_i)$

Les 280 noeuds correspondent aux aéroports. Leurs 8 attributs représentent des informations sur le trafic (p.ex. augmentation/diminution des annulations, des vols). Les 1 206 arêtes correspondent à l'existence de vols entre les aéroports. Le troisième jeu de données est lié au suivi de l'activité aquacole en Indonésie à partir d'une série d'images satellitaires. Le processus mis en place pour ces données ainsi que les résultats obtenus sont détaillés dans [Che18], ils ne seront pas présentés dans ce chapitre.

Plusieurs jeux de données synthétiques ont aussi été générés afin d'étudier plus en détail l'influence de certains paramètres des données sur le passage à l'échelle. Nous avons plus particulièrement étudié l'influence du nombre de noeuds par temps, du nombre d'attributs par noeud, du nombre d'arêtes, et du nombre de temps. Le générateur de graphes dynamiques attribués développé prend donc en paramètre ces différents éléments. Il génère le nombre de noeuds demandé, puis affecte à chacun de ces noeuds un certain nombre d'attributs aux différents temps selon une distribution uniforme. Les arêtes sont ensuite générées à partir de paires de noeuds choisies selon aussi une distribution uniforme.

4.3.3 Analyse qualitative

Une analyse qualitative des motifs dans les jeux de données réels a été effectuée. Dans le jeu de données $DBLP$, cette analyse a mis en avant certains comportements récurrents de groupes d'auteurs. Dans le jeu de données sur le trafic aérien, les motifs ont confirmé et précisé l'influence de plusieurs cyclones sur le trafic aérien aux Etats-Unis. Afin d'illustrer cela, deux exemples de motifs sont présentés dans la suite.

Le premier motif est issu du jeu de données *DBLP* et présente les publications faites par deux co-auteurs pendant la période de 1990 à 2009. Il est formellement décrit par

$$\Big(< \quad (Henry\ Tirri : KDD, ICML \mid Petri\ Myllymaki : KDD, ICML) \\ (Henry\ Tirri : KDD, IntellDtAnal \mid Petri\ Myllymaki : KDD, IntellDtAnal) \\ (Henry\ Tirri : ECMLPKDD \mid Petri\ Myllymaki : ECAI) >, \quad \{[90-93], [94-97]\} \Big)$$

Ce motif montre un comportement récurrent de Henry Tirri et Petri Myllymaki (2 répétitions). Ils publient tout d'abord ensemble dans *KDD* et *ICML*, puis publient dans *KDD* et *Intelligent Data Analysis*, et pour finir *ECML/PKDD* et *ECAI* indépendamment. Au-delà de l'intérêt potentiel de ce motif, il est intéressant de souligner que ce type d'évolution n'aurait pas pu être extrait par les approches existantes de part les domaines de motifs très spécifiques considérés.

Le deuxième motif présente l'impact des cyclones Katrina, Irene et Ophelia sur une partie du trafic aérien des Etats-Unis (3 récurrences). Ces cyclones ont eu lieu sur la côte Est. Ils ont donc eu un impact important dans cette zone en terme d'annulations et de retards, mais pas uniquement. Ils ont aussi eu un impact indirect sur plusieurs aéroports de la côte Ouest. Comme attendu, cet impact est proportionnel à la puissance du cyclone. La figure 4.4 présente le graphe associé à ce motif.

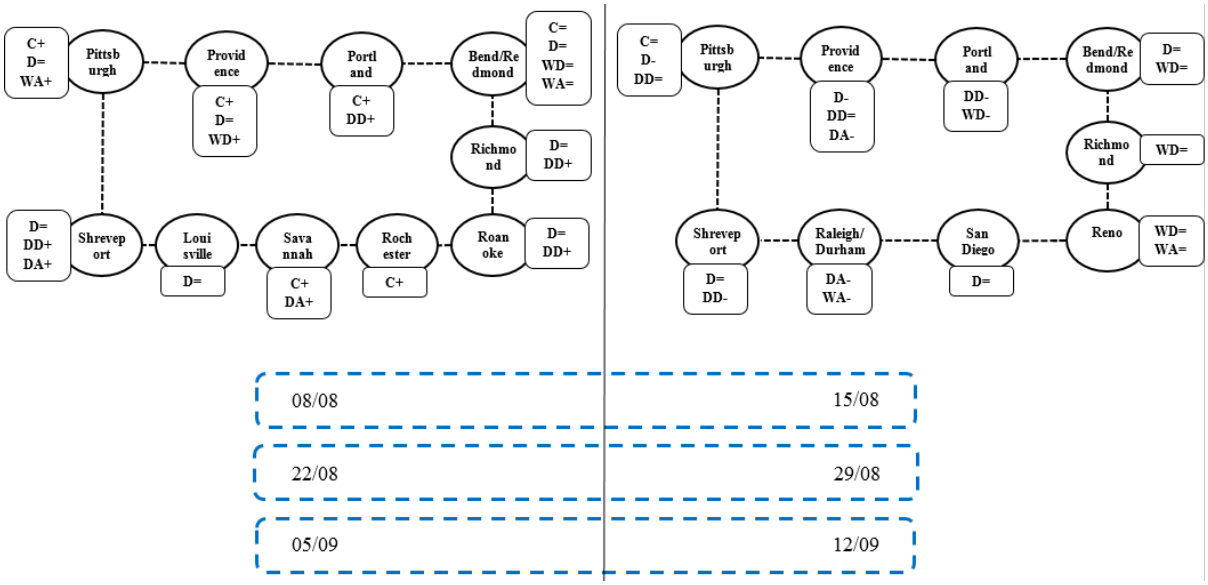


FIGURE 4.4 – Exemple de motif présentant l'impact de trois cyclones sur les vols aux Etats-Unis. *C : annulations, D : vols redirigés, DD : délai moyen au départ, DA : délai moyen à l'arrivée, WD : temps d'attente au sol au départ, WA : temps d'attente au sol à l'arrivée.*

4.3.4 Analyse quantitative

La figure 4.5 présente les temps d'exécution et le nombre de solutions obtenus pour 12 jeux de données synthétiques comportant un nombre croissant de sommets et d'arêtes (avec 50 attributs en moyenne et 8 dates). Comme le montrent ces graphiques, l'algorithme reste relativement

efficace lors de l'analyse d'une séquence composée de 20 000 sommets par graphe et par date, avec un seuil très petit de fréquence ($minsup = 2$).

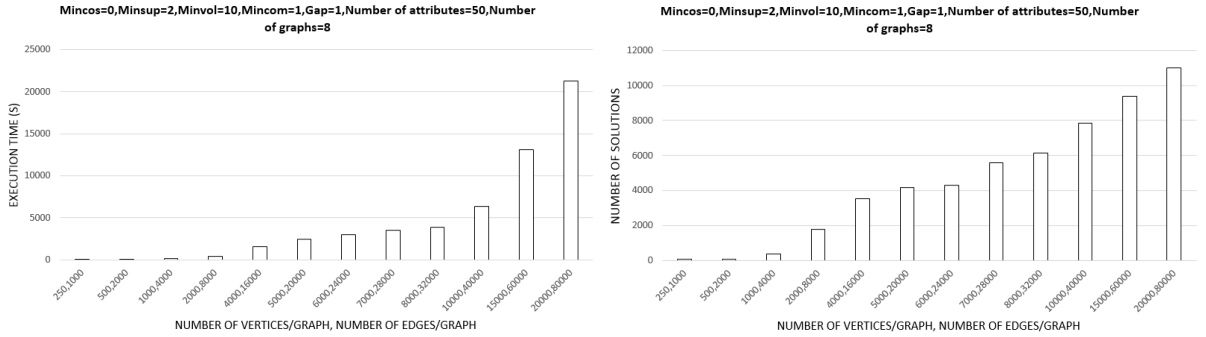


FIGURE 4.5 – Temps d'exécution et nombre de solutions en fonction du nombre de noeuds et d'arêtes (données synthétiques)

La figure 4.6 montre que le temps d'exécution et la mémoire maximale consommée (tout comme le nombre de solutions) augmentent de manière exponentielle en fonction du nombre de temps (de graphes dans la séquence). Cet impact est très important en raison de la stratégie utilisée par l'algorithme. En effet, le nombre d'intersections effectué est exponentiel dans le nombre de temps (une intersection pour chaque combinaison de temps de taille supérieure à *minsup*). Toutefois, même si le passage à l'échelle est plus difficile pour des graphes dynamiques avec beaucoup de temps, l'algorithme proposé permet de traiter de plus grands graphes que les algorithmes de la littérature, tout en trouvant des solutions plus générales.

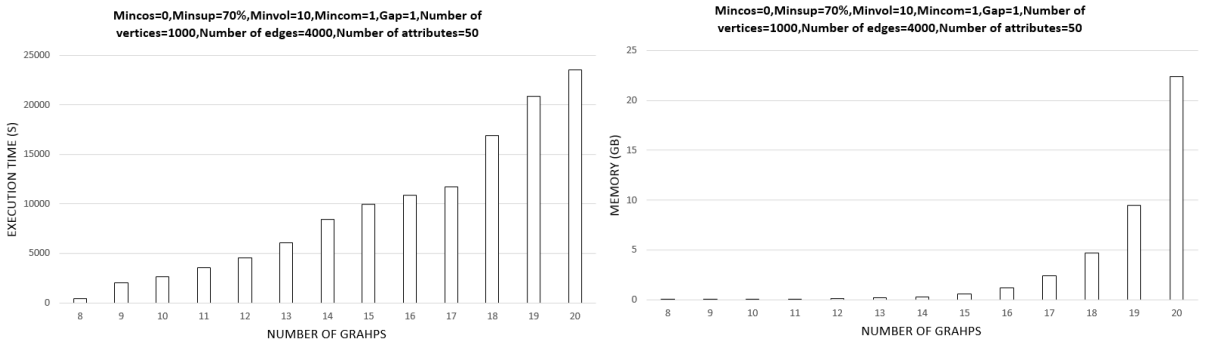


FIGURE 4.6 – Temps d'exécution et mémoire utilisée en fonction du nombre de temps (données synthétiques)

Le nombre d'attributs impacte beaucoup moins les performances que le nombre de temps. Comme le montre la figure 4.7, le temps d'exécution augmente linéairement avec le nombre d'attributs. L'algorithme peut ainsi traiter plus de 1000 attributs par noeud, ce qui permet de modéliser et d'analyser des données très riches en informations.

Des expérimentations ont aussi été conduites afin d'étudier l'impact des contraintes sur les performances. La contrainte qui a le plus d'impact est le seuil de fréquence *minsup*. En effet, ce seuil conditionne directement le nombre d'intersections effectuées. Seules les intersections de plus *minsup* graphes sont calculées pour respecter la contrainte de fréquence minimale. Les autres contraintes ont un impact plus ou moins important en fonction de la densité du jeu de

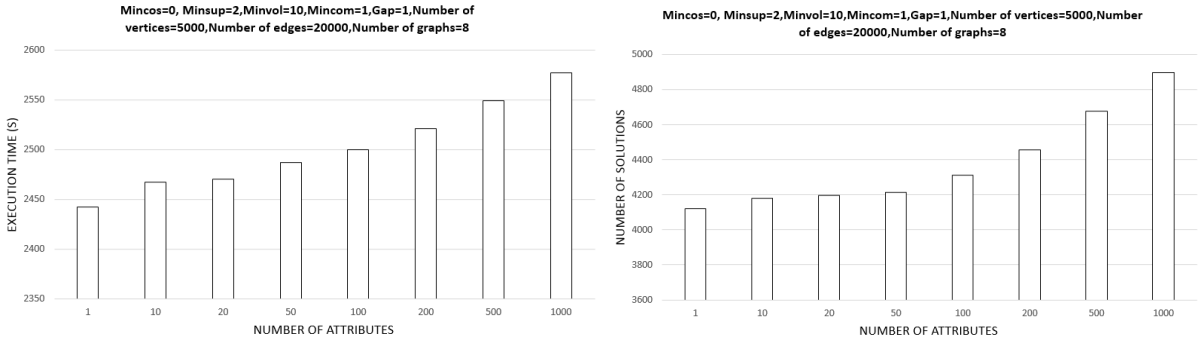


FIGURE 4.7 – Temps d'exécution et nombre de solutions en fonction du nombre d'attributs (données synthétiques)

données (de la taille des composantes connexes notamment). Dans tous les cas, ils améliorent les performances et diminuent le nombre de solutions.

Quatrième partie

Conclusion et perspectives

1

Synthèse

Le travail présenté dans cette habilitation à diriger des recherches montre l'évolution de mes thématiques de recherche en accord avec les problématiques locales et les verrous identifiés dans la communauté. Je suis ainsi parti de mes travaux de thèse sur l'extraction d'*itemsets* fréquents dans des données "*benchmarks*", pour aller vers de l'extraction de motifs dans des données spatio-temporelles réelles en interaction avec des experts du domaine. La complexité de ces données réelles a mis en avant les limites des approches existantes, et nous a permis de proposer des solutions originales et génériques à la communauté. De plus, la proximité des experts nous a amené à nous intéresser à l'intérêt des motifs extraits ainsi qu'à leur restitution. J'ai ainsi travaillé à l'élaboration des domaines de motifs et d'algorithmes de plus en plus complexes : des co-localisations aux graphes dynamiques attribués, en passant par des séquences intégrant le voisinage.

Le chapitre 1 a décrit un premier travail s'intéressant à la découverte et la restitution de motifs spatiaux (des co-localisations) plus pertinents pour les experts. L'intégration de la connaissance du domaine dans l'analyse, associée à une visualisation adaptée des résultats, est essentielle pour fournir aux experts des connaissances utiles et interprétables. L'étude des outils et des pratiques des experts, nous a amené à proposer de nouvelles contraintes et une nouvelle approche de visualisation. Deux types de contraintes ont été proposées : des contraintes de type "SIG" définies manuellement par les experts, et des contraintes dérivées des modèles mathématiques issus de la littérature du domaine. Comme l'ont montré les expérimentations réalisées dans le cadre du suivi de l'érosion, ces contraintes permettent de centrer l'analyse sur des motifs d'intérêt et d'avoir une extraction plus efficace. La visualisation cartographique proposée, et basée sur un *clustering* des instances des motifs, permet quant à elle de mettre en avant de manière intuitive la distribution spatiale des solutions. Ces propositions ont été intégrées dans des outils SIG et validées avec des experts du domaine.

Le chapitre 2 a présenté un nouveau domaine de motifs séquentiels intégrant le voisinage. Ainsi, les dimensions spatiales et temporelles sont en partie prises en compte dans ce travail. Ce domaine de motifs est utilisé pour analyser les évolutions dans une zone (fixe) tout en prenant en compte l'environnement proche. Le cadre théorique des motifs séquentiels et l'algorithme $PrefixSpan$ [PHMA$^+$01] sont étendus pour intégrer cette nouvelle dimension. La stratégie de $PrefixSpan$ basée sur des projections successives est modifiée pour inclure aussi les informations du voisinage. De nouvelles extensions sont ajoutées à la liste des extensions possibles des préfixes. Une mesure d'intérêt est aussi proposée en post-traitement afin de filtrer les motifs les plus contredits. Ces contributions ont été utilisées, avec les experts, pour analyser les dynamiques d'une maladie vectorielle (la dengue) dans une ville et pour étudier la pollution de plusieurs

rivières. L'interface de visualisation développée a permis de discuter plus simplement des résultats avec les experts.

Le chapitre 3 s'est focalisé sur l'évolution d'objets spatiaux caractérisés par des dynamiques complexes (p.ex. déplacement, apparition/disparition, fusion/division). Un graphe orienté acyclique attribué (ou *a-DAG*) est utilisé pour modéliser ces dynamiques. Dans ce contexte, les évolutions sont étudiées via un nouveau domaine de motifs appelé chemin pondéré. Ces motifs correspondent à des chemins fréquents et non-redondants dans le *a-DAG*. L'algorithme développé pour les extraire permet d'analyser efficacement des graphes de taille relativement importante (200 000 noeuds, 600 000 arêtes et 7.5 attributs par noeud en moyenne). Sa stratégie globale est similaire à $PrefixSpan$. Toutefois, le principe des extensions est très différent (extensions, partielles et complètes, par des *itemsets*), tout comme les projections effectuées et la structure de données utilisée (graphe de motifs). Ces contributions ont notamment été utilisées pour analyser une série d'images satellitaires liées à l'érosion des sols en Nouvelle-Calédonie. Un processus complet d'extraction de connaissances a été mis en place et intégré dans la plate-forme KNIME. Les résultats obtenus ont souligné l'évolution de l'érosion en lien avec la végétation, la pente et l'activité minière. Les expérimentations sur d'autres types de données ont démontré la généricité de l'approche développée.

Le chapitre 4 intègre de manière plus complète les dimensions spatiales et temporelles. Pour cela, ce travail s'appuie sur une représentation des données sous forme de graphe dynamique attribué. Les relations spatiales peuvent ainsi être finement représentées par un graphe, et leurs évolutions dans le temps par l'aspect dynamique du graphe. Ce graphe est utilisé pour extraire les évolutions récurrentes d'objets liés spatialement. Ces évolutions constituent un nouveau domaine de motifs correspondant à des sous-séquences de composantes connexes sous-contraintes (notamment structurelles et temporelles). Une stratégie originale et incrémentale a été proposée. Elle s'appuie sur plusieurs propriétés des intersections de sous-graphes connexes. Cette stratégie permet de traiter des graphes plus grands que dans la littérature, mais surtout beaucoup plus riches en informations (1000 attributs). Cette approche a été utilisée dans différents contextes applicatifs (réseau de co-auteurs, trafic aérien et aquaculture). Les motifs obtenus ont montré l'intérêt de ce domaine de motifs par rapport aux autres existants et sa généricité.

Discussion D'un point de vue plus général, la modélisation des données sous-forme de graphe dynamique attribué est la plus générale et la plus complète. Elle permet de capturer un grand nombre d'interactions spatiales et temporelles, tout en permettant d'intégrer l'ensemble des informations associées à chaque objet. Elle a aussi l'avantage d'être générique avec un grand nombre de domaines d'application potentiels (bien au delà des données spatio-temporelles). Les problèmes de recherches de co-localisations et de motifs spatio-séquentiels peuvent être directement reformulés dans ce cadre. La recherche des évolutions fréquentes dans une série d'images satellitaires peut l'être aussi à condition d'intégrer des contraintes supplémentaires lors de l'analyse afin de capturer les dynamiques de fusion/division et les déplacements potentiels.

Toutefois, extraire des motifs dans ce type de graphes est particulièrement complexe. L'approche développée actuellement, même si elle est plus générale que les contributions existantes, reste assez spécifique. Elle ne permet pas d'extraire les motifs fréquents, seulement les motifs récurrents (i.e. des motifs apparaissant au même endroit mais à des dates différentes). Il n'est donc pas possible de l'utiliser pour extraire des motifs spatio-séquentiels ou des évolutions fréquentes (la localisation des objets n'entrant pas directement en jeu dans le calcul de la fréquence). Il s'agit d'ailleurs d'une des perspectives de ces travaux qui sera développée dans le chapitre suivant.

Les domaines de motifs proposés restent donc encore complémentaires. Le travail sur les mo-

tifs spatio-séquentiels permet d'extraire les évolutions fréquentes des attributs de zones fixes, en prenant en compte les attributs des zones voisines. Le travail sur les chemins pondérés fréquents permet d'extraire des évolutions fréquentes d'objets spatiaux ayant des dynamiques plus complexes, mais n'intègre pas les informations des objets à proximité. Le travail sur les sous-graphes attribués permet de capturer assez finement les interactions spatiales et temporelles (même si les arêtes ne sont pas directement considérées), mais se limite à rechercher des récurrences locales et non des évolutions fréquentes globalement. Bien entendu, il est totalement envisageable de combiner toutes ces contributions pour extraire des motifs plus généraux. Toutefois, l'analyse d'un graphe dynamique attribué pose des problèmes de passage à l'échelle pour lesquels un simple assemblage de techniques existantes risque de ne pas être suffisant.

A noter que la contribution sur l'extraction de motifs guidée par des modèles du domaine est une approche générique et applicable à d'autres domaines de motifs. Elle est présentée dans le cadre du travail sur les co-localisations pour des questions de présentation, mais elle a aussi été utilisée lors de l'analyse de séries d'images satellitaires avec les chemins pondérés fréquents. En effet, cette approche s'applique, pour l'instant, uniquement à des *itemsets*. Or, les *itemsets* sont présents dans les motifs séquentiels et les graphes attribués. Il est donc possible d'utiliser ces contraintes dérivées de modèles dans ces contextes pour filtrer certaines solutions (p.ex. certaines extensions de motifs séquentiels).

2

Perspectives

Les perspectives à ces travaux sont multiples. Il s'agit de perspectives directes de nos derniers travaux, mais aussi de problématiques qui m'intéressent et que je souhaiterais approfondir. Certaines sont d'ailleurs actuellement au centre de travaux de recherche actuels au sein de l'équipe. Elles se situent dans le domaine de l'extraction de motifs mais aussi dans d'autres domaines telle que la classification. Elles s'articulent autour de quatre thèmes :

- la fouille de graphes (amélioration du passage à l'échelle et extraction de motifs plus généraux)
- l'extraction de motifs au sens large (intégration de la discrétisation et prise en compte de la connaissance du domaine)
- la classification (prise en compte du bruit et des valeurs manquantes)
- les applications (renforcement des liens avec le privé et élargissement à d'autres problématiques)

Fouille de graphes. Dans les derniers travaux présentés, nous avons étudié l'extraction de motifs dans des graphes complexes : les graphes dynamiques attribués. Cette représentation a l'avantage d'être très générale et de pouvoir représenter un grand nombre de données (p.ex. images, vidéos, réseaux sociaux, informations géographiques, molécules, etc.). Toutefois, leur analyse est difficile. Le simple fait de savoir si deux graphes sont isomorphes est un problème pour lequel on ne connaît pas d'algorithmes en temps polynomial dans le cas général. Rechercher toutes les corrélations possibles dans ces graphes, en un temps raisonnable, est donc un défi. L'approche proposée en chapitre 4 permet de traiter plus de données et d'analyser des corrélations plus complexes que dans la littérature. Toutefois, deux points pourraient encore être améliorés : le passage à l'échelle de l'algorithme et le domaine de motifs.

L'une des limites de l'algorithme développé est la consommation de mémoire. A chaque itération, toutes les solutions sont conservées en mémoire afin d'être potentiellement étendues lors de l'itération suivante (i.e. lors du traitement du temps suivant). Comme le montre la figure 4.6, la mémoire utilisée explose lorsque les données sont associées à plus de 20 dates (plus de 20 Go de mémoire). Face à ce problème, une solution est de mettre en place une structure de données plus optimisée pour stocker les motifs (l'implémentation actuelle utilise les structures de données de la librairie *Boost Graph*). La mise en place d'une telle structure de données a généralement un impact sur l'algorithme, et nécessite donc de l'adapter. Une autre solution est de changer la stratégie de l'algorithme. Certaines stratégies ont la particularité de limiter la mémoire utilisée. Les stratégies effectuant un parcours en profondeur de l'espace recherche ont par exemple cette

particularité. L'algorithme actuel est d'ailleurs en cours d'adaptation afin de suivre un parcours en profondeur, et les premiers résultats montrent déjà une amélioration des performances. Au delà de l'aspect purement algorithmique, l'extraction peut aussi être améliorée en tirant partie des nouvelles architectures logicielles et matérielles développées ces dernières années. L'intérêt croissant autour du *big data* et du *cloud computing* a donné naissance à de nouveaux concepts, architectures et solutions logicielles tels que le paradigme *MapReduce*, *Hadoop* ou *Apache Spark*. La démocratisation des appareils mobiles et des objets connectés, conjointement avec une utilisation accrue de "l'intelligence artificielle", amène aussi de nouvelles avancées matérielles telles que le processeur *Edge TPU* de Google, le *Kirin 980* de Huawei ou le *LightOn* de la *start-up* du même nom. L'exploitation de ces technologies permettrait d'optimiser ou de distribuer certaines opérations. L'algorithme proposé, de par sa stratégie incrémentale basée sur des intersections et des extensions en parallèle des motifs, semble particulièrement adapté à ce type d'architectures.

Les motifs récurrents constituent un domaine de motifs plus général que ceux proposés dans la littérature pour analyser des graphes dynamiques attribués. Toutefois, ils correspondent encore à des régularités très locales. Si deux évolutions identiques sont associées à des noeuds différents, elles seront considérées comme des motifs différents. Il serait donc intéressant de généraliser encore ce domaine de motifs pour extraire ces évolutions identiques indépendamment des noeuds d'apparition. On obtiendrait ainsi des sous-graphes fréquents et non simplement récurrents. Ce domaine de motifs est plus complexe à extraire et soulève donc des défis en terme de passage à l'échelle. Les solutions évoquées précédemment seront donc encore plus d'actualité avec ce nouveau domaine de motifs. Ces motifs plus généraux peuvent aussi être obtenus par post-traitement, évitant ainsi la définition d'un nouveau domaine de motifs et d'un nouvel algorithme. Une option est d'exploiter un algorithme de *clustering* pour regrouper les motifs similaires et ainsi retourner des informations plus générales à l'expert du domaine. Dans la littérature, des mesures de similarités ont été proposées pour certains domaines de motifs telles que les séquences. Il faudrait donc les adapter aux graphes dynamiques attribués. Ainsi, les motifs satisfaisant une similarité minimale pourraient être regroupés et résumés. Pour cela, l'algorithme proposé dans [AHSZ11] serait une solution intéressante. Il impose uniquement une borne inférieure sur la similarité entre chaque objet et le représentant du *cluster* associé, et supporte les chevauchements de *clusters*. Différents problèmes restent encore à étudier à ce niveau, comme le passage à l'échelle de l'ensemble du processus (extraction de motifs et *clustering*) ou la caractérisation de l'information contenue dans chaque cluster.

Extraction de motifs au sens large. Dans un processus d'extraction et de gestion des connaissances, l'extraction des motifs est précédée par différentes opérations de sélection et de pré-traitement des données. Parmi toutes ces opérations, la discrétisation est une opération classique lorsque l'on veut extraire des motifs. En effet, la majeure partie des algorithmes d'extraction de motifs nécessitent d'avoir en entrée des données catégorielles (encore appelées qualitatives ou binaires). Lorsque l'on a des données numériques, il est donc nécessaire de les discrétiser, i.e. de les transformer en données catégorielles (généralement des intervalles disjoints). Différentes méthodes existent (p.ex. méthode des quantiles ou méthodes des amplitudes) [DKS95], mais dans tous les cas, l'impact de cette opération sur les résultats est très important. Plusieurs travaux ont étudié cette problématique [SA96a, KKN11, GS18]. Ils discrétisent les données "à la volée" pendant l'extraction ou traitent directement des données numériques (en utilisant des mesures de distances). Toutefois, ils souffrent encore d'un certain nombre de limites : perte d'informations, analyse mono-dimensionnelle, passage à l'échelle ou non intégration des utilisateurs. Je souhaiterais notamment étudier cette problématique dans le cadre de données spatio-temporelles où les

notions de hiérarchie et de granularité sont particulièrement importantes (avec des corrélations potentiellement à différents niveaux).

La pertinence des motifs extraits est également un enjeu important. Un grand nombre de mesures et de contraintes ont été proposées dans la littérature afin d'avoir des motifs plus intéressants d'un point de vue statistique, mais aussi du point de vue des experts du domaine d'application. Pour cela, une approche classique consiste à intégrer la connaissance des experts sous forme de règles afin de guider l'extraction. Toutefois, ce type d'approche nécessite une forte implication des experts pour définir et ajuster ces règles. Dans notre travail, nous avons proposé d'exploiter les modèles mathématiques existants dans la littérature du domaine. Par exemple, il existe de nombreux modèles mathématiques visant à modéliser et simuler l'érosion des sols. Ces modèles sont définis par les experts à partir de leur connaissance du phénomène étudié. Ils prennent en compte certaines interactions fines entre variables, mais restent limités dans le nombre de variables considérées. Notre idée était de tirer partie de cette connaissance et d'en dériver des contraintes utilisées pendant l'extraction de motifs pour améliorer la pertinence de l'analyse. Notre travail s'est focalisé sur des modèles "statiques" applicables à des *itemsets*. Il ne prend pas en compte la dynamique temporelle des phénomènes étudiés. Or, des modèles intégrant cette dynamique existent. Ils se présentent souvent sous la forme de systèmes d'équations différentielles. Le modèle *SIR* (*Susceptible*, *Infectious*, *Recovered*) utilisé pour modéliser l'évolution d'une maladie infectieuse en est un exemple. Il serait donc particulièrement intéressant d'exploiter aussi ce type de modèles pour améliorer l'extraction de motifs (p.ex. séquences ou graphes dynamiques). Au delà d'une simple utilisation de ces modèles orientés simulation comme contrainte, je souhaiterais étudier un réel couplage entre ces techniques et celles développées en fouille de données. Un tel couplage permettrait par exemple de paramétrer de manière semi-supervisée ces modèles théoriques, ou de les enrichir, en croisant données observées et données simulées. Ce travail initierait aussi des interactions entre informaticiens et physiciens. En effet, une des thématiques étudiées par les physiciens de l'ISEA est le développement de modèles mathématiques intégrant une composante chaotique pour simuler différents phénomènes réels (p.ex. lasers).

Classification. Au delà de l'extraction de motifs, je souhaiterais m'investir dans une autre tâche classique en analyse de données : l'apprentissage supervisé. Mes travaux sur l'extraction de motifs pourraient être un point de départ. En effet, l'un des résultats récents dans le domaine de l'extraction de motifs est l'intérêt de ces modèles locaux en tant que descripteurs dans les problèmes de classification supervisés. Plusieurs méthodes de classification supervisées basées sur des règles d'association, appelées CBA (*Classification Based on Associations*), ont été développées dans le cadre de données binaires, ce qui a entraîné un gain significatif d'informations [ML98, LDR01, AZ04, Tha07, GSFB12, CGSF$^+$13, ZN14]. Je souhaiterais plus particulièrement étudier cette problématique dans le cadre des séries temporelles multivariées (*multivariate time series*). L'analyse de séries temporelles est une problématique au centre de nombreux travaux actuellement [Mue14, TL17]. L'une des principales difficultés est de pouvoir croiser et prédire l'évolution dans le temps de plusieurs variables numériques. La majeure partie des travaux vise à prédire le futur à partir de la valeur courante de la série. Des travaux récents ont montré qu'étudier la forme de certaines parties de la série pouvait donner de meilleurs résultats et permettait d'aboutir à la construction de règles. Ces travaux se sont pour l'instant limités à l'étude d'une variable et à des prédictions ponctuelles en fonction des motifs détectés. Il serait intéressant de généraliser cette notion de règles à des séries temporelles multivariées et de combiner ces règles pour prédire les futures valeurs.

Lorsque l'on doit traiter des données réelles, une autre problématique importante est la présence de bruit dans les données et l'absence ponctuelle de certaines valeurs [ZW04]. Ces problèmes peuvent venir de la précision des capteurs ou des erreurs commises par les experts lors de la saisie des informations. Les performances des méthodes d'apprentissage sont généralement fortement impactées par ces deux types d'anomalies. Une grande partie des solutions reste dans le cadre du filtrage, du nettoyage des données ou de l'interpolation. Elles mettent aussi souvent l'accent sur le bruit associé à la phase d'apprentissage (p.ex. bruit de classes) [FV14]. Peu d'études ont étudié les problèmes des erreurs affectant les attributs (mesures), alors que ce type d'anomalies est plus nuisible. Dans ce contexte, une première approche pourrait être de définir et d'utiliser des motifs tolérants aux bruits dans le cadre des séries temporelles multivariées.

Un autre défi important en cours d'étude dans le cadre de l'analyse les séries temporelles est de prédire le plus "tôt" possible (*early prediction*), i.e. prédire les évolutions (et les risques) dès que possible [LCTP15]. Dans un tel contexte, la précision des classificateurs n'est pas le seul critère important. La "précocité" par rapport à la dimension temporelle doit également être considérée, comme la stabilité de la précision.

Applications. Je souhaiterais aussi élargir mon champ applicatif au delà des problématiques environnementales. Un grand nombre de domaines sont possibles mais plusieurs applications m'intéressent plus particulièrement en raison de leur intérêt local et de la possibilité de créer des partenariats avec des entreprises. Les domaines en question sont l'agriculture, l'industrie et la santé.

La dernière thèse que je co-encadre se place par exemple dans le contexte de l'aquaculture. Elle vise à développer de nouvelles méthodes pour analyser et comprendre les facteurs influençant la réussite des élevages de crevettes. Elle est en partenariat avec l'IFREMER, et les acteurs de la filière aquacole en Nouvelle-Calédonie (le Groupement des Fermes Aquacoles et la SOciété des Producteurs Aquacoles Calédoniens). Ce travail s'appuie pour cela sur des séries temporelles multivariées bruitées et hétérogènes issues de différentes sources de données telles que les données d'élevage, les données de qualité et les données météorologiques. D'un point de vue méthodologique, on s'intéresse dans ce travail à des problématiques de classification (supervisée et non supervisée) avec pour principal verrou l'imprécision des mesures et les valeurs manquantes.

Au niveau industriel, un partenariat est en cours de mise en place avec une entreprise dans le secteur de la production d'énergies renouvelables (solaire et éolienne). L'objectif de ce partenariat est de développer des méthodes permettant de prédire à partir des séries de données collectées en continu sur les sites de production. Deux problématiques intéressent plus particulièrement l'entreprise : la prévision des pannes des éoliennes et la prévision de la production solaire. Bien que ces applications semblent différentes, elles partagent le même type de données, à savoir des séries temporelles multivariées avec des valeurs numériques et nominales. Pour l'instant, ce partenariat s'est appuyé sur deux stages et un projet tuteuré orientés davantage développement logiciel et bases de données, mais l'objectif à terme est d'obtenir un financement de thèse pour approfondir cette problématique. Des discussions sont également en cours avec une entreprise du secteur minier sur des questions de ressources (et non de suivi environnemental).

Le domaine de la santé est aussi un domaine producteur de données avec des questions liées au suivi des patients et à l'optimisation des processus internes. Ces deux dernières années, la situation a beaucoup évolué à ce niveau en Nouvelle-Calédonie. Les installations étaient relativement anciennes, certaines datant de la seconde guerre mondiale. Un programme de modernisation a donc été mis en place par le gouvernement et le secteur privé. Ce programme a abouti à l'ouverture d'un médipôle en 2016 et aux regroupements des différentes cliniques dans un même site

en 2018, sous le nom de clinique Kuindo-Magnin. Ces nouvelles installations bénéficient des dernières technologies au niveau médical avec un suivi des patients informatisé et des instruments connectés. Une grande quantité de données est en cours de production actuellement avec des besoins importants en terme de valorisations et des spécificités locales à prendre en compte.

Bibliographie

[AA99] Gennady Andrienko and Natalia Andrienko. Knowledge-based visualization to support spatial data mining. In *International Symposium on Intelligent Data Analysis*, pages 149–160. Springer, 1999.

[AA16] Berkay Aydin and Rafal A Angryk. A graph-based approach to spatiotemporal event sequence mining. In *Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1090–1097. IEEE Computer Society, 2016.

[ABK+07] Luis Otavio Alvares, Vania Bogorny, Bart Kuijpers, Jose Antonio Fernandes de Macedo, Bart Moelans, and Alejandro Vaisman. A model for enriching trajectories with semantic geographical information. In *Proceedings of the ACM international symposium on Advances in geographic information systems*, page 22. ACM, 2007.

[ABP11] Alberto Apostolico, Manuel Barbares, and Cinzia Pizzi. Speedup for a periodic subgraph miner. *Information Processing Letters*, 111(11) :521–523, 2011.

[AFGY02] Jay Ayres, Jason Flannick, Johannes Gehrke, and Tomi Yiu. Sequential pattern mining using a bitmap representation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 429–435. ACM, 2002.

[AHSZ11] Mohammad Al Hasan, Saeed Salem, and Mohammed J Zaki. Simclus : an effective algorithm for clustering with a lower bound on similarity. *Knowledge And Information Systems*, 28(3) :665–685, 2011.

[AK15] Rezwan Ahmed and George Karypis. Algorithms for mining the coevolving relational motifs in dynamic networks. *ACM Transactions on Knowledge Discovery from Data*, 10(1) :4, 2015.

[AKK18] Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. Spatio-temporal data mining : A survey of problems and methods. *ACM Computing Surveys (CSUR)*, 51(4) :83, 2018.

[AMGACO+16] Niusvel Acosta-Mendoza, Andrés Gago-Alonso, Jesús Ariel Carrasco-Ochoa, José Francisco Martínez-Trinidad, and José Eladio Medina-Pagola. Improving graph-based image classification by using emerging patterns as attributes. *Engineering Applications of Artificial Intelligence*, 50 :215–225, 2016.

[AMGAMP12] Niusvel Acosta-Mendoza, Andrés Gago-Alonso, and José E Medina-Pagola. Frequent approximate subgraphs as features for graph-based image classification. *Knowledge-Based Systems*, 27 :381–392, 2012.

[Ant08] Cláudia Antunes. An ontology-based framework for mining patterns in the presence of background knowledge. In *Proceedings of the 1st International Conference on Advanced Intelligence (ICAI)*, pages 163–168, 2008.

[AS+94] Rakesh Agarwal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, pages 487–499, 1994.

[AS95] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 3–14. IEEE Computer Society, 1995.

[ASG13] Hamed Abdelhaq, Christian Sengstock, and Michael Gertz. Eventweet : Online localized event detection from twitter. *Proceedings of the VLDB Endowment*, 6(12) :1326–1329, 2013.

[Ath05] James Atherton. Watershed Assessment for Healthy Reefs and Fisheries. Technical Report 679, 2005.

[AZ04] Maria-Luiza Antonie and Osmar R Zaïane. An associative classifier based on positive and negative rules. In *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 64–69. ACM, 2004.

[AZC01] Maria-Luiza Antonie, Osmar R Zaiane, and Alexandru Coman. Application of data mining techniques for medical image classification. In *Proceedings of the Second International Conference on Multimedia Data Mining*, pages 94–101. Springer-Verlag, 2001.

[Azé03] Jérôme Azé. Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances. *Revue d'intelligence artificielle*, 17(1-3) :171–182, 2003.

[Bar11] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1-3) :1–101, 2011.

[Bay98] Roberto J Bayardo Jr. Efficiently Mining Long Patterns from Databases. In Laura M Haas and Ashutosh Tiwary, editors, *SIGMOD Conference*, pages 85–93. ACM Press, 1998.

[BB02] Christian Borgelt and Michael R Berthold. Mining molecular fragments : Finding relevant substructures of molecules. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 51–58. IEEE Computer Society, 2002.

[BBBG09] Michele Berlingerio, Francesco Bonchi, Björn Bringmann, and Aristides Gionis. Mining graph evolution rules. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 115–130. Springer, 2009.

[BBR03] Jean-François Boulicaut, Artur Bykowski, and Christophe Rigotti. Free-sets : a condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery*, 7(1) :5–22, 2003.

[BCD+07] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. KNIME : The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007.

[BCG01] Douglas Burdick, Manuel Calimlim, and Johannes Gehrke. Mafia : A maximal frequent itemset algorithm for transactional databases. In *Proceedings of the International Conference on Data Engineering*, pages 443–452. IEEE Computer Society, 2001.

[BDF+03] VG Blinova, DA Dobrynin, VK Finn, Sergei O Kuznetsov, and ES Pankratova. Toxicology analysis by means of the jsm-method. *Bioinformatics*, 19(10) :1201–1207, 2003.

[BHPG11] Albert Bifet, Geoff Holmes, Bernhard Pfahringer, and Ricard Gavaldà. Mining frequent closed graphs on evolving data streams. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 591–599. ACM, 2011.

[BL10] Enrico Bertini and Denis Lalanne. Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *SIGKDD Exploration Newsletter*, 11(2) :9–18, 2010.

[BM93] S. Beucher and F. Meyer. The morphological approach to segmentation : the watershed transformation. Mathematical morphology in image processing. *Optical Engineering*, 34 :433–481, 1993.

[BMB04] Christian Borgelt, Thorsten Meinl, and Michael R Berthold. Advanced pruning strategies to speed up mining closed molecular fragments. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 5, pages 4565–4570. IEEE Computer Society, 2004.

[BMBH95] A Bannari, D Morin, F Bonn, and AR Huete. A review of vegetation indices. *Remote sensing reviews*, 13(1-2) :95–120, 1995.

[BMML15] Peter A Burrough, Rachael McDonnell, Rachael A McDonnell, and Christopher D Lloyd. *Principles of geographical information systems*. Oxford University Press, 2015.

[BMS11] Petko Bogdanov, Misael Mongiovì, and Ambuj K Singh. Mining heavy subgraphs in time-evolving networks. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 81–90. IEEE Computer Society, 2011.

[BN08] Björn Bringmann and Siegfried Nijssen. What is frequent in a single graph ? In *Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 858–863, 2008.

[BRdA+14] Vania Bogorny, Chiara Renso, Artur Ribeiro de Aquino, Fernando de Lucca Siqueira, and Luis Otavio Alvares. Constant–a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS*, 18(1) :66–88, 2014.

[BTD17] Berna Bakır Batu, Tuğba Taşkaya Temizel, and H Şebnem Düzgün. A non-parametric algorithm for discovering triggering patterns of spatio-temporal event types. *IEEE Transactions on Knowledge and Data Engineering*, 29(12) :2629–2642, 2017.

[BVd+06] Vania Bogorny, João Francisco Valiati, Sandro da Silva Camargo, Paulo Martins Engel, Bart Kuijpers, and Luis Otávio Alvares. Mining Maximal Generalized Frequent Geographic Patterns with Knowledge Constraints. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 813–817. IEEE Computer Society, 2006.

[CBRB08] Loïc Cerf, Jérémy Besson, Céline Robardet, and Jean-François Boulicaut. Data-peeler : Constraint-based closed pattern mining in n-ary relations. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 37–48. SIAM, 2008.

[CCL05] Alain Casali, Rosine Cicchetti, and Lotfi Lakhal. Essential patterns : A perfect cover of frequent patterns. In *Proceedings of the International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, pages 428–437. Springer, 2005.

[Cel15] Mete Celik. Partial spatio-temporal co-occurrence pattern mining. *Knowledge And Information Systems*, 44(1) :27–49, 2015.

[CGSF+13] Loïc Cerf, Dominique Gay, Nazha Selmaoui-Folcher, Bruno Crémilleux, and Jean-François Boulicaut. Parameter-free classification in multi-class imbalanced data sets. *Data & Knowledge Engineering*, 87 :109–129, 2013.

[CH94] Diane J. Cook and Lawrence B. Holder. Substructure discovery using minimum description length and background knowledge. *Journal of Artificial Intelligence Research*, 1 :231–255, 1994.

[Che10] Jinlin Chen. An updown directed acyclic graph approach for sequential pattern mining. *IEEE Transactions on Knowledge and Data Engineering*, 22(7) :913–928, 2010.

[Che18] Zhi Cheng. *Mining recurrent petterns in a dynamic attributed graph. Application to aquaculture Pond Monitoring by satellite images.* PhD thesis, Noumea, University of New Caledonia, 2018.

[CI09] Emmanuel Christophe and Jordi Inglada. Open source remote sensing : Increasing the usability of cutting-edge algorithms. *IEEE Geoscience and Remote Sensing Newsletter*, 35(5) :9–15, 2009.

[CKK04] Yen-liang Chen, Hung-pin Kao, and Ming-tat Ko. Mining DAG Patterns from DAG Databases. *Advances in Web-Age Information Management*, pages 579–588, 2004.

[CMC05] Huiping Cao, Nikos Mamoulis, and David W Cheung. Mining Frequent Spatio-Temporal Sequential Patterns. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 82–89. IEEE Computer Society, 2005.

[CNB09] Loïc Cerf, Tran Bao Nhan Nguyen, and Jean-François Boulicaut. Discovering relevant cross-graph cliques in dynamic networks. In *International symposium on methodologies for intelligent systems*, pages 513–522. Springer, 2009.

[Cre93] Noel Cressie. *Statistics for Spatial Data.* Wiley-Interscience, rev sub edition, 1993.

[CSRS06] Mete Celik, Shashi Shekhar, James P Rogers, and James A Shine. Sustained Emerging Spatio-Temporal Co-occurrence Pattern Mining : A Summary of Results. In *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 106–115, 2006.

[CSRS08] Mete Celik, Shashi Shekhar, James P Rogers, and James A Shine. Mixed-drove spatiotemporal co-occurrence pattern mining. *IEEE Transactions on Knowledge and Data Engineering*, 20(10) :1322–1335, 2008.

[CVDVM16] Alberto Coletta, Victor Van Der Veen, and Federico Maggi. Droydseuss : A mobile banking trojan tracker (short paper). In *International Conference on Financial Cryptography and Data Security*, pages 250–259. Springer, 2016.

[dCB+11] Líliam César de Castro Medeiros, César Augusto Rodrigues Castilho, Cynthia Braga, Wayner Vieira de Souza, Leda Regis, and Antonio Miguel Vieira Monteiro. Modeling the dynamic transmission of dengue fever : investigating disease persistence. *PLoS neglected tropical diseases*, 5(1) :e942, January 2011.

[DFJS16] Romain Deville, Elisa Fromont, Baptiste Jeudy, and Christine Solnon. Grima : a grid mining algorithm for bag-of-grid-based classification. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 132–142. Springer, 2016.

[DJK11] Balázs Dezső, Alpár Jüttner, and Péter Kovács. Lemon–an open source c++ graph template library. *Electronic Notes in Theoretical Computer Science*, 264(5) :23–45, 2011.

[DKS95] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. In *Machine Learning Proceedings 1995*, pages 194–202. Elsevier, 1995.

[DLLS15] Brahim Douar, Michel Liquiere, Chiraz Latiri, and Yahya Slimani. Lc-mine : a framework for frequent subgraph mining with local consistency techniques. *Knowledge And Information Systems*, 44(1) :1–25, 2015.

[DLSL09] Bertrand De Longueville, Robin S Smith, and Gianluca Luraschi. Omg, from here, i can see the flames ! : a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 international workshop on location based social networks*, pages 73–80. ACM, 2009.

[DPDR01] Qin Ding, William Perrizo, Qiang Ding, and Amalendu Roy. On mining satellite and other remotely sensed images. In *Data Mining and Knowledge Discovery*, 2001.

[DPRB12] Elise Desmier, Marc Plantevit, Céline Robardet, and Jean-François Boulicaut. Cohesive co-evolution patterns in dynamic attributed graphs. In *Proceedings of the International Conference on Discovery Science (DS)*, pages 110–124. Springer, 2012.

[DPRB13] Elise Desmier, Marc Plantevit, Céline Robardet, and Jean-François Boulicaut. Trend mining in dynamic attributed graphs. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 654–669. 2013.

[ECGFS10] Ashraf Elsayed, Frans Coenen, Marta García-Fiñana, and Vanessa Sluming. Region of interest based image categorization. In *Proceedings of the International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, pages 239–250. Springer, 2010.

[EKSX96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 226–231, 1996.

[ES02] Martin Erwig and Markus Schneider. Spatio-temporal predicates. *IEEE Transactions on Knowledge and Data Engineering*, (4) :881–901, 2002.

[FAL+13] Anna Fariha, Chowdhury Farhan Ahmed, Carson Kai-Sang Leung, SM Abdullah, and Longbing Cao. Mining frequent patterns from human interactions in meetings using directed acyclic graphs. In *Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 38–49. Springer, 2013.

[FB74] Raphael A. Finkel and Jon Louis Bentley. Quad trees a data structure for retrieval on composite keys. *Acta informatica*, 4(1) :1–9, 1974.

[FB07] Mathias Fiedler and Christian Borgelt. Support Computation for Mining Frequent Subgraphs in a Single Graph. In *Mining and Learning with Graphs*, 2007.

[FDMP09] Frédéric Flouvat, Fabien De Marchi, and Jean-Marc Petit. The iZi project : easy prototyping of interesting pattern mining algorithms. In *Advanced Techniques for Data Mining and Knowledge Discovery*, LNCS, pages 1–15. Springer-Verlag, 2009.

[FFT12] Basura Fernando, Elisa Fromont, and Tinne Tuytelaars. Effective use of frequent itemset mining for image classification. In *European Conference on Computer Vision*, pages 214–227. Springer, 2012.

[FGCOMT14] Marisol Flores-Garrido, Jesús Ariel Carrasco-Ochoa, and JF Martínez-Trinidad. Mining maximal frequent patterns in a single graph using inexact matching. *Knowledge-Based Systems*, 66 :166–177, 2014.

[FK14] James H Faghmous and Vipin Kumar. Spatio-temporal data mining for climate data : Advances, challenges, and opportunities. In *Data Mining and Knowledge Discovery for Big Data*, pages 83–116. Springer, 2014.

[FNVSDSF15] Frédéric Flouvat, Jean-François N'guyen Van Soc, Elise Desmier, and Nazha Selmaoui-Folcher. Domain-driven co-location mining. *Geoinformatica*, 19(1) :147–183, 2015.

[FSKS10] Mutsumi Fukuzaki, Mio Seki, Hisashi Kashima, and Jun Sese. Finding itemset-sharing patterns in a large itemset-associated graph. In *Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 147–159, 2010.

[FV14] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise : a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5) :845–869, 2014.

[FVGCT14] Philippe Fournier-Viger, Antonio Gomariz, Manuel Campos, and Rincy Thomas. Fast vertical mining of sequential patterns using co-occurrence information. In *Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 40–52. Springer, 2014.

[FVGŠH14] Philippe Fournier-Viger, Antonio Gomariz, Michal Šebek, and Martin Hlosta. Vgen : fast vertical mining of sequential generator patterns. In *Proceedings of the International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, pages 476–488. Springer, 2014.

[FVLK+17] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage Uday Kiran, Yun Sing Koh, and Rincy Thomas. A survey of sequential pattern mining. *Data Science and Pattern Recognition*, 1(1) :54–77, 2017.

[FVWT13] Philippe Fournier-Viger, Cheng-Wei Wu, and Vincent S Tseng. Mining maximal sequential patterns without candidate maintenance. In *International Conference on Advanced Data Mining and Applications*, pages 169–180. Springer, 2013.

[GB17] Andrew Gilbert and Richard Bowden. Image and video mining through online learning. *Computer Vision and Image Understanding*, 158 :72–84, 2017.

[GCMG13] Antonio Gomariz, Manuel Campos, Roque Marin, and Bart Goethals. Clasp : An efficient algorithm for mining frequent closed sequences. In *Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 50–61. Springer, 2013.

[GG89] Michael F Goodchild and Sucharita Gopal. *The accuracy of spatial databases*. CRC Press, 1989.

[GNPP07] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory pattern mining. In Pavel Berkhin, Rich Caruana, and Xindong Wu, editors, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 330–339. ACM, 2007.

[GP08] Fosca Giannotti and Dino Pedreschi, editors. *Mobility, Data Mining and Privacy - Geographic Knowledge Discovery*. Springer, 2008.

[GRS99] Minos N Garofalakis, Rajeev Rastogi, and Kyuseok Shim. Spirit : Sequential pattern mining with regular expression constraints. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, volume 99, pages 7–10, 1999.

[Grü07] Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

[GS10] Stephan Günnemann and Thomas Seidl. Subgraph Mining on Directed and Weighted Graphs. In *Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 133–146, 2010.

[GS18] Arnaud Giacometti and Arnaud Soulet. Dense neighborhood pattern sampling in numerical data. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 756–764. SIAM, 2018.

[GSFB12] Dominique Gay, Nazha Selmaoui-Folcher, and Jean-François Boulicaut. Application-independent feature construction based on almost-closedness properties. *Knowledge And Information Systems*, 30(1) :87–111, 2012.

[GSV06] Ehud Gudes, Solomon Eyal Shimony, and Natalia Vanetik. Discovering Frequent Graph Patterns Using Disjoint Paths. *IEEE Transactions on Knowledge and Data Engineering*, 18(11) :1441–1456, 2006.

[GYC07] Michael F Goodchild, May Yuan, and Thomas J Cova. Towards a general theory of geographic representation in gis. *International journal of geographical information science*, 21(3) :239–260, 2007.

[GZ03] Gösta Grahne and Jianfei Zhu. Efficiently Using Prefix-trees in Mining Frequent Itemsets. In Roberto J Bayardo Jr. and Mohammed Javeed Zaki, editors, *FIMI*, volume 90 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2003.

[HC+09] Lawrence B Holder, Diane J Cook, et al. Learning patterns in the dynamics of biological networks. In *Proceedings of the ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 977–986. ACM, 2009.

[HK01] Rie Honda and Osamu Konishi. Temporal rule discovery for time-series satellite images and integration with rdb. In *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, pages 204–215. Springer, 2001.

[HPL15] Chih-Chieh Hung, Wen-Chih Peng, and Wang-Chien Lee. Clustering and aggregating clues of trajectories for mining trajectory patterns and routes. *The International Journal on Very Large Data Bases*, 24(2) :169–192, 2015.

[HPT12] Phan Nhat Hai, Pascal Poncelet, and Maguelonne Teisseire. Get_move : an efficient and unifying spatio-temporal pattern mining algorithm for moving objects. In *International Symposium on Intelligent Data Analysis*, pages 276–288. Springer, 2012.

[HPY00] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *ACM sigmod record*, volume 29, pages 1–12. ACM, 2000.

[HSL17] Sajal Halder, Md Samiullah, and Young-Koo Lee. Supergraph based periodic pattern mining in dynamic social networks. *Expert Systems with Applications*, 72 :430–442, 2017.

[HSX04] Yan Huang, Shashi Shekhar, and Hui Xiong. Discovering colocation patterns from spatial data sets : a general approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(12) :1472–1485, 2004.

[HZZ08] Yan Huang, Liqin Zhang, and Pusheng Zhang. A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Transactions on Knowledge and Data Engineering*, 20(4) :433–448, 2008.

[IW08] Akihiro Inokuchi and Takashi Washio. A fast method to mine frequent subsequences from graph sequence data. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 303–312. IEEE Computer Society, 2008.

[IW10] Akihiro Inokuchi and Takashi Washio. Mining frequent graph sequence patterns induced by vertices. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 466–477. SIAM, 2010.

[IWM00] Akihiro Inokuchi, Takashi Washio, and Hiroshi Motoda. An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data. In *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, volume 1910, pages 13–23. Springer, 2000.

[Jac12] Paul Jaccard. The distribution of the flora in the alpine zone. *New phytologist*, 11(2) :37–50, 1912.

[JC09] Chuntao Jiang and Frans Coenen. Graph-based image classification by weighting scheme. In *Applications and Innovations in Intelligent Systems XVI*, pages 63–76. Springer, 2009.

[JCZ10] Chuntao Jiang, Frans Coenen, and Michele Zito. Frequent sub-graph mining on edge weighted graphs. In *Proceedings of the International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, pages 77–88. Springer, 2010.

[JCZ13] Chuntao Jiang, Frans Coenen, and Michele Zito. A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review*, 28(01) :75–105, 2013.

[JMB+11] Andreea Julea, Nicolas Méger, Philippe Bolon, Christophe Rigotti, Marie-Pierre Doin, Cécile Lasserre, Emmanuel Trouvé, and Vasile N Lazarescu. Unsupervised spatiotemporal mining of satellite image time series using grouped frequent sequential patterns. *IEEE Transactions on Geoscience and Remote Sensing*, 49(4) :1417–1430, 2011.

[JXWT09] Xing Jiang, Hui Xiong, Chen Wang, and Ah-Hwee Tan. Mining globally distributed frequent subgraphs in a single labeled graph. *Data & Knowledge Engineering*, 68(10) :1034–1058, 2009.

[JZH11] Yi Jia, Jintao Zhang, and Jun Huan. An efficient graph-mining method for complicated and noisy data with real-world applications. *Knowledge And Information Systems*, 28(2) :423–447, 2011.

[KHA98] Krzysztof Koperski, Jiawei Han, and Junas Adhikary. Mining knowledge in geographical data. *Communications of ACM (accepted)*, 1998.

[KK01] Michihiro Kuramochi and George Karypis. Frequent subgraph discovery. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 313–320. IEEE Computer Society, 2001.

[KK05] Michihiro Kuramochi and George Karypis. Finding Frequent Patterns in a Large Sparse Graph*. *Data Mining and Knowledge Discovery*, 11(3) :243–271, September 2005.

[KKN11] Mehdi Kaytoue, Sergei O Kuznetsov, and Amedeo Napoli. Revisiting numerical pattern mining with formal concept analysis. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1342, 2011.

[KO02] Sergei O Kuznetsov and Sergei A Obiedkov. Comparing performance of algorithms for generating concept lattices. *Journal of Experimental & Theoretical Artificial Intelligence*, 14(2-3) :189–216, 2002.

[KPPR15] Mehdi Kaytoue, Yoann Pitarch, Marc Plantevit, and Céline Robardet. What effects topological changes in dynamic graphs ? *Social Network Analysis and Mining*, 5(1) :55, 2015.

[KPWD03] Hye-Chung Kum, Jian Pei, Wei Wang, and Dean Duncan. Approxmap : Approximate mining of consensus sequential patterns. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 311–315. SIAM, 2003.

[KW03] Michael Kaufmann and Dorothea Wagner. Drawing graphs : methods and models, 2003.

[KYW10] Arijit Khan, Xifeng Yan, and Kun-Lung Wu. Towards proximity pattern mining in large graphs. *Proceedings of the 2010 international conference on Management of data - SIGMOD '10*, page 867, 2010.

[LBW08] Mayank Lahiri and Tanya Y Berger-Wolf. Mining periodic behavior in dynamic social networks. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 373–382. IEEE Computer Society, 2008.

[LC05] Congnan Luo and Soon M Chung. Efficient mining of maximal sequential patterns using multiple samples. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 415–426. SIAM, 2005.

[LCTP15] Yu-Feng Lin, Hsuan-Hsu Chen, Vincent S Tseng, and Jian Pei. Reliable early classification on multivariate time series with numerical and categorical attributes. In *Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 199–211. Springer, 2015.

[LDR01] Jinyan Li, Guozhu Dong, and Kotagiri Ramamohanarao. Making use of the most expressive jumping emerging patterns for classification. *Knowledge And Information Systems*, 3(2) :131–145, 2001.

[LG12] Olivier Lézoray and Leo Grady. *Image processing and analysis with graphs : theory and practice*. CRC Press, 2012.

[LHJ+11] Zhenhui Li, Jiawei Han, Ming Ji, Lu-An Tang, Yintao Yu, Bolin Ding, Jae-Gil Lee, and Roland Kays. Movemine : Mining moving object data for discovery of animal movement patterns. *ACM Transactions on Intelligent Systems and Technology*, 2(4) :37, 2011.

[LHW07] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. Trajectory clustering : a partition-and-group framework. In *Proceedings of the ACM SIGMOD international conference on Management of data*, pages 593–604. ACM, 2007.

[LIC08] Carson Kai-Sang Leung, Pourang Irani, and Christopher L Carmichael. WiFIsViz : Effective Visualization of Frequent Itemsets. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 875–880. IEEE Computer Society, 2008.

[LK14] Jure Leskovec and Andrej Krevl. SNAP Datasets : Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[LKF05] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time : densification laws, shrinking diameters and possible explanations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 177–187. ACM, 2005.

[LKL07] David Lo, Siau-Cheng Khoo, and Chao Liu. Efficient mining of iterative patterns for software specification discovery. In *Proceedings of the ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 460–469. ACM, 2007.

[LKL08] David Lo, Siau-Cheng Khoo, and Jinyan Li. Mining and ranking generators of sequential patterns. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 553–564. SIAM, 2008.

[Llo82] Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2) :129–137, 1982.

[LLSvdH15] Yao Li, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Mid-level deep pattern mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 971–980, 2015.

[LMBM04] Dengsheng Lu, P Mausel, E Brondizio, and Emilio Moran. Change detection techniques. *International Journal of Remote Sensing*, 25(12) :2365–2401, 2004.

[LMFC12] Hoang Thanh Lam, Fabian Moerchen, Dmitriy Fradkin, and Toon Calders. Mining compressing sequential patterns. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 319–330, 2012.

[LMFC14] Hoang Thanh Lam, Fabian Mörchen, Dmitriy Fradkin, and Toon Calders. Mining compressing sequential patterns. *Statistical Analysis and Data Mining : The ASA Data Science Journal*, 7(1) :34–52, 2014.

[LN89] L.J. Lane and M.A. Nearing. USDA - Water Erosion Prediction Project : Hillslope Profile Model Documentation. 1989.

[LXS17] Xiaohan Liao, Cunjin Xue, and Fenzhen Su. Tree-based approach for exploring marine spatial patterns with raster datasets. *PloS one*, 12(5) :e0177438, 2017.

[LZ05] Benjamin Livshits and Thomas Zimmermann. Dynamine : finding common error patterns by mining software revision histories. In *ACM SIGSOFT Software Engineering Notes*, volume 30, pages 296–305. ACM, 2005.

[LZC+11] Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xie Xing. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1010–1018. ACM, 2011.

[LZY10] Shirong Li, Shijie Zhang, and Jiong Yang. DESSIN : mining dense subgraph patterns in a single graph. In *Proceedings of the International Conference on Scientific and Statistical Database Management (SSDBM)*, pages 178–195, 2010.

[MBTP04] Behrouz Minaei-Bidgoli, Pang-Ning Tan, and William F Punch. Mining interesting contrast rules for a web-based educational system. In *Proceedings of the International Conference on Machine Learning and Applications (ICMLA)*, pages 320–327. IEEE Computer Society, 2004.

[MCK+04] Nikos Mamoulis, Huiping Cao, George Kollios, Marios Hadjieleftheriou, Yufei Tao, and David W Cheung. Mining, indexing, and querying historical spatiotemporal data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 236, 2004.

[MCP98] Florent Masseglia, Fabienne Cathala, and Pascal Poncelet. The PSP Approach for Mining Sequential Patterns. In Jan M Zytkow and Mohamed Quafafou, editors, *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, volume 1510 of *Lecture Notes in Computer Science*, pages 176–184. Springer, 1998.

[MCRE09] Flavia Moser, Recep Colak, Arash Rafiey, and Martin Ester. Mining Cohesive Patterns from Graphs with Feature Vectors. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 593–604, 2009.

[MCUP04] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10) :761–767, 2004.

[ME10] Nizar R Mabroukeh and Christie I Ezeife. A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys (CSUR)*, 43(1) :3, 2010.

[MHMW00] Takashi Matsuda, Tadashi Horiuchi, Hiroshi Motoda, and Takashi Washio. Extension of Graph-Based Induction for General Graph Structured Data. In *Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 420–431, 2000.

[ML98] Bing Liu Wynne Hsu Yiming Ma and Bing Liu. Integrating classification and association rule mining. In *Proceedings of the international conference on knowledge discovery and data mining*, 1998.

[ML08] David Marsan and Olivier Lengliné. Extending earthquakes' reach through cascading. *Science*, 319(5866) :1076–1079, 2008.

[MOO09] Yuuki Miyoshi, Tomonobu Ozaki, and Takenao Ohkawa. Frequent Pattern Discovery from a Single Graph with Quantitative Itemsets. *2009 IEEE International Conference on Data Mining Workshops*, pages 527–532, December 2009.

[Mor01] R.P.C Morgan. A simple approach to soil loss prediction : a revised Morgan-Morgan-Finney model. *Catena*, 44(4) :305–322, July 2001.

[MR11] Muhammad Muzammal and Rajeev Raman. Mining sequential patterns from probabilistic databases. In *Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 210–221. Springer, 2011.

[MR13] Carl H Mooney and John F Roddick. Sequential pattern mining–approaches and algorithms. *ACM Computing Surveys (CSUR)*, 45(2) :19, 2013.

[MRP15] Nicolas Méger, Christophe Rigotti, and Catherine Pothier. Swap randomization of bases of sequences for mining satellite image times series. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 190–205. Springer, 2015.

[MSSR10] Pradeep Mohan, Shashi Shekhar, James A Shine, and James P Rogers. Cascading Spatio-temporal Pattern Discovery : A Summary of Results. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 327–338, 2010.

[MSSR12] Pradeep Mohan, Shashi Shekhar, James A Shine, and James P Rogers. Cascading spatio-temporal pattern discovery. *IEEE Transactions on Knowledge and Data Engineering*, 24(11) :1977–1992, 2012.

[MTV97a] Heikki Mannila, Hannu Toivonen, and A Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3) :259–289, 1997.

[MTV97b] Heikki Mannila, Hannu Toivonen, and A Inkeri Verkamo. Levelwise Search and Borders of Theories in Knowledge Discovery. *Data Mining and Knowledge Discovery*, 1(3) :241–258, 1997.

[Mue14] Abdullah Mueen. Time series motif discovery : dimensions and applications. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 4(2) :152–159, 2014.

[MWW+15] Yan Ma, Haiping Wu, Lizhe Wang, Bormin Huang, Rajiv Ranjan, Albert Zomaya, and Wei Jie. Remote sensing big data computing : Challenges and opportunities. *Future Generation Computer Systems*, 51 :47–60, 2015.

[NCLY07] Vincent Ng, Stephen Chan, Derek Lau, and Cheung Man Ying. Incremental mining for temporal association rules for crime pattern discoveries. In *Proceedings of the 18th conference on Australasian database*, pages 123–132. Australian Computer Society, 2007.

[NK04] Siegfried Nijssen and Joost N. Kok. A quickstart in frequent structure mining can make a difference. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 647–652, New York, NY, USA, 2004. ACM.

[NNP+09] Tung Thanh Nguyen, Hoan Anh Nguyen, Nam H. Pham, Jafar M. Al-Kofahi, and Tien N. Nguyen. Graph-based mining of multiple object usage patterns. In *Proceedings of the the Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on The Foundations of Software Engineering*, ESEC/FSE '09, pages 383–392, New York, NY, USA, 2009. ACM.

[NTU+07] Sebastian Nowozin, Koji Tsuda, Takeaki Uno, Taku Kudo, and Gokhan BakIr. Weighted substructure mining for image analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE Computer Society, 2007.

[OA10] Bahadir Ozdemir and Selim Aksoy. Image classification using subgraph histogram representation. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 1112–1115. IEEE Computer Society, 2010.

[OE11] Tomonobu Ozaki and Minoru Etoh. Closed and maximal subgraph mining in internally and externally weighted graph databases. In *IEEE Workshops of the International Conference on Advanced Information Networking and Applications*, pages 626–631. IEEE Computer Society, 2011.

[OO09] Tomonobu Ozaki and Takenao Ohkawa. Discovery of correlated sequential subgraphs from a sequence of graphs. In *International Conference on Advanced Data Mining and Applications*, pages 265–276. Springer, 2009.

[OPS04] Salvatore Orlando, Raffaele Perego, and Claudio Silvestri. A new algorithm for gap constrained sequence mining. In *Proceedings of the 2004 ACM symposium on Applied computing*, pages 540–547. ACM, 2004.

[PAA13] Karthik Ganesan Pillai, Rafal A Angryk, and Berkay Aydin. A filter-and-refine approach to mine spatiotemporal co-occurrences. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 104–113. ACM, 2013.

[Par94] Dillon Pariente. *Estimation, modélisation et langage de déclaration et de manipulation de champs spatiaux continus.* PhD thesis, Lyon, INSA, 1994.

[Pat10] Dhaval Patel. Interval-orientation patterns in spatio-temporal databases. In *International Conference on Database and Expert Systems Applications*, pages 416–431. Springer, 2010.

[PBTL99] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering Frequent Closed Itemsets for Association Rules. In *Proceedings of the International Conference on Database Theory (ICDT)*, pages 398–416. Springer, 1999.

[PCL+05] Marc Plantevit, Yeow Wei Choong, Anne Laurent, Dominique Laurent, and Maguelonne Teisseire. M 2 sp : Mining sequential patterns among several dimensions. In *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, pages 205–216. Springer, 2005.

[PFSSF17] Claude Pasquier, Frédéric Flouvat, Jérémy Sanhes, and Nazha Selmaoui-Folcher. Attributed graph mining in the presence of automorphism. *Knowledge And Information Systems*, 50(2) :569–584, 2017.

[PGMF10] François Petitjean, Pierre Gançarski, Florent Masseglia, and Germain Forestier. Analysing satellite image time series by means of pattern mining. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 45–52. Springer, 2010.

[Pha13] NhatHai Phan. *Mining Object Movement Patterns from Trajectory Data. (Une approche unifiée pour extraire des motifs à partir de données de trajectoires)*. PhD thesis, Montpellier 2 University, France, 2013.

[PHMA+01] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu. Prefixspan : Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, page 215. IEEE Computer Society, 2001.

[PIG12] François Petitjean, Jordi Inglada, and Pierre Gançarski. Satellite image time series analysis under time warping. *IEEE Transactions on Geoscience and Remote Sensing*, 50(8) :3081–3095, 2012.

[PIV+15] Yoann Pitarch, Dino Ienco, Elodie Vintrou, Agnès Bégué, Anne Laurent, Pascal Poncelet, Michel Sala, and Maguelonne Teisseire. Spatio-temporal data classification through multidimensional sequential patterns : Application to crop mapping in complex landscape. *Engineering Applications of Artificial Intelligence*, 37 :91–102, 2015.

[PJFD13] Adriana Prado, Baptiste Jeudy, Élisa Fromont, and Fabien Diot. Mining spatiotemporal patterns in dynamic plane graphs. *Intelligent Data Analysis*, 17(1) :71–92, 2013.

[PM00] Dan Pelleg and Andrew W Moore. X-means : Extending K-means with Efficient Estimation of the Number of Clusters. In Pat Langley, editor, *Proceedings of the International Conference on Machine Learning (ICML)*, pages 727–734. Morgan Kaufmann, 2000.

[PS15] Vyoma Patel and GJ Sahani. Image classification using frequent itemset mining. *International Journal of Computer Applications*, 121(15), 2015.

[PSFSF16] Claude Pasquier, Jérémy Sanhes, Frédéric Flouvat, and Nazha Selmaoui-Folcher. Frequent pattern mining in attributed trees : algorithms and applications. *Knowledge And Information Systems*, 46(3) :491–514, 2016.

[PSTV10] Luca Paolino, Monica Sebillo, Genoveffa Tortora, and Giuliana Vitiello. Integrating discrete and continuous data in an opengeospatial-compliant specification. *Transactions in GIS*, 14(6) :731–753, 2010.

[QHH09] Feng Qian, Qinming He, and Jiangfeng He. Mining Spread Patterns of Spatiotemporal Co-occurrences over Zones. In Osvaldo Gervasi, David Taniar, Beniamino Murgante, Antonio Laganà, Youngsong Mun, and Marina L Gavrilova, editors, *Proceedings of the International Conference on Computational Science and Its Applications*, volume 5593 of *Lecture Notes in Computer Science*, pages 677–692. Springer, 2009.

[RdAC+13] Luciana Alvim S Romani, Ana Maria H de Avila, Daniel YT Chino, Jurandir Zullo, Richard Chbeir, Caetano Traina, and Agma JM Traina. A new time series mining approach applied to multitemporal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1) :140–150, 2013.

[RFDT15] Konstantinos Rematas, Basura Fernando, Frank Dellaert, and Tinne Tuytelaars. Dataset fingerprints : Exploring image collections through data mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4867–4875, 2015.

[RPT10] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2) :9, 2010.

[SA96a] Ramakrishnan Srikant and Rakesh Agrawal. Mining quantitative association rules in large relational tables. In *ACM Sigmod Record*, volume 25, pages 1–12. ACM, 1996.

[SA96b] Ramakrishnan Srikant and Rakesh Agrawal. Mining sequential patterns : Generalizations and performance improvements. In *International Conference on Extending Database Technology*, pages 1–17. Springer, 1996.

[SC78] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1) :43–49, 1978.

[SDC+13] Chengnian Sun, Jing Du, Ning Chen, Siau-Cheng Khoo, and Ye Yang. Mining explicit rules for software process evaluation. In *Proceedings of the International Conference on Software and System Process*, pages 118–125. ACM, 2013.

[SEKM11] Shashi Shekhar, Michael R Evans, James M Kang, and Pradeep Mohan. Identifying patterns in spatial information : A survey of methods. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 1(3) :193–214, 2011.

[SFF11] Nazha Selmaoui-Folcher and Frédéric Flouvat. How to use classical tree mining algorithms to find complex spatio-temporal patterns ? In *International Conference on Database and Expert Systems Applications*, pages 107–117. Springer, 2011.

[SGE12] Saurabh Singh, Abhinav Gupta, and Alexei A Efros. Unsupervised discovery of mid-level discriminative patches. In *Computer Vision–ECCV 2012*, pages 73–86. Springer, 2012.

[SH01] Shashi Shekhar and Yan Huang. Discovering Spatial Co-location Patterns : A Summary of Results. In Christian S Jensen, Markus Schneider, Bernhard Seeger, and Vassilis J Tsotras, editors, *Proceedings of the International Symposium on Spatial and Temporal Databases (SSTD)*, volume 2121 of *Lecture Notes in Computer Science*, pages 236–256. Springer, 2001.

[SJA+15] Shashi Shekhar, Zhe Jiang, Reem Y Ali, Emre Eftelioglu, Xun Tang, Venkata Gunturi, and Xun Zhou. Spatiotemporal data mining : a computational perspective. *ISPRS International Journal of Geo-Information*, 4(4) :2306–2338, 2015.

[SK02] Masakazu Seno and George Karypis. Slpminer : An algorithm for finding frequent sequential patterns using length-decreasing support constraint. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 418–425. IEEE Computer Society, 2002.

[SK14] Mirka Saarela and Tommi Kärkkäinen. Discovering gender-specific knowledge from finnish basic education using pisa scale indices. In *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*. International Educational Data Mining Society (IEDMS), 2014.

[SLL01] Jeremy G Siek, Lie-Quan Lee, and Andrew Lumsdaine. *The Boost Graph Library : User Guide and Reference Manual*. Pearson Education, 2001.

[SMJZ12] Arlei Silva, Wagner Meira Jr, and Mohammed J Zaki. Mining attribute-structure correlated patterns in large attributed graphs. *Proceedings of the VLDB Endowment*, 5(5) :466–477, 2012.

[SVCS09] Daniel Sánchez, MA Vila, L Cerda, and José-Maria Serrano. Association rules applied to credit card fraud detection. *Expert systems with applications*, 36(2) :3630–3640, 2009.

[SVN10] Laszlo Szathmary, Petko Valtchev, and Amedeo Napoli. Generating rare association rules using the minimal rare itemsets family. *International Journal Software Informatics*, 4(3), 2010.

[SW11] Robert Sedgewick and Kevin Wayne. *Algorithms, 4th Edition.* Addison-Wesley, 2011.

[T+06] Pang-Ning Tan et al. *Introduction to data mining.* Pearson Education India, 2006.

[TG01] Ilias Tsoukatos and Dimitrios Gunopulos. Efficient Mining of Spatiotemporal Patterns. In Christian S Jensen, Markus Schneider, Bernhard Seeger, and Vassilis J Tsotras, editors, *Proceedings of the International Symposium on Spatial and Temporal Databases (SSTD)*, volume 2121 of *Lecture Notes in Computer Science*, pages 425–442. Springer, 2001.

[Tha07] Fadi Thabtah. A review of associative classification mining. *The Knowledge Engineering Review*, 22(1) :37–65, 2007.

[TKS02] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 32–41. ACM, 2002.

[TL17] Sahar Torkamani and Volker Lohweg. Survey on time series motif discovery. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 7(2) :e1199, 2017.

[Tob70] Waldo Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2) :234–240, 1970.

[TRS02] Alexandre Termier, Marie-Christine Rousset, and Michèle Sebag. TreeFinder : a First Step towards XML Data Mining. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 450–457. IEEE Computer Society, 2002.

[TTN+07] Alexandre Termier, Yoshinori Tamada, Kazuyuki Numata, Seiya Imoto, Takashi Washio, Tomoyuki Higushi, and Tomoyuki Higuchi. DigDag, a first algorithm to mine closed frequent embedded sub-DAGs. *MLG*, pages 1–5, 2007.

[TV12] Nikolaj Tatti and Jilles Vreeken. The long and the short of it : summarising event sequences with serial episodes. In *Proceedings of the ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 462–470. ACM, 2012.

[UAUA03] Takeaki Uno, Tatsuya Asai, Yuzo Uchida, and Hiroki Arimura. LCM : An Efficient Algorithm for Enumerating Frequent Closed Item Sets. In Roberto J Bayardo Jr. and Mohammed Javeed Zaki, editors, *FIMI*, volume 90 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2003.

[UAUA04] Takeaki Uno, Tatsuya Asai, Yuzo Uchida, and Hiroki Arimura. An Efficient Algorithm for Enumerating Closed Patterns in Transaction Databases. In Einoshin Suzuki and Setsuo Arikawa, editors, *Proceedings of the International Conference on Discovery Science (DS)*, volume 3245 of *Lecture Notes in Computer Science*, pages 16–31. Springer, 2004.

[VAN07] Florian Verhein and Ghazi Al-Naymat. Fast Mining of Complex Spatial Colocation Patterns Using GLIMIT. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, pages 679–684. IEEE Computer Society, 2007.

[Vap63] Vladimir Vapnik. Pattern recognition using generalized portrait method. *Automation and remote control*, 24 :774–780, 1963.

[VCJ14] Winn Voravuthikunchai, Bruno Crémilleux, and Frédéric Jurie. Histograms of pattern sets for image classification and object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 224–231, 2014.

[VGS02] Natalia Vanetik, Ehud Gudes, and Solomon Eyal Shimony. Computing frequent graph patterns from semistructured data. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 458–465. IEEE Computer Society, 2002.

[VVLS11] Jilles Vreeken, Matthijs Van Leeuwen, and Arno Siebes. Krimp : mining itemsets that compress. *Data Mining and Knowledge Discovery*, 23(1) :169–214, 2011.

[WD04] Michael F Worboys and Matt Duckham. *GIS : a computing perspective*. CRC press, 2004.

[WDL+17] Guojun Wu, Yichen Ding, Yanhua Li, Jie Bao, Yu Zheng, and Jun Luo. Mining spatio-temporal reachable regions over massive trajectory data. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, pages 1283–1294. IEEE Computer Society, 2017.

[WDW+08] Tobias Werth, Alexander Dreweke, Marc Wörlein, Ingrid Fischer, and Michael Philippsen. DAGMA : Mining Directed Acyclic Graphs. *IADIS European Conference on Data Mining*, 2008.

[WF94] Stanley Wasserman and Katherine Faust. *Social network analysis : Methods and applications*, volume 8. Cambridge university press, 1994.

[WHL05] Junmei Wang, Wynne Hsu, and Mong-Li Lee. Mining Generalized Spatio-Temporal Patterns. In Lizhu Zhou, Beng Chin Ooi, and Xiaofeng Meng, editors, *Proceedings of the International Conference on Database Systems for Advanced Applications (DASFAA)*, volume 3453 of *Lecture Notes in Computer Science*, pages 649–661. Springer, 2005.

[WHL07] Jianyong Wang, Jiawei Han, and Chun Li. Frequent closed sequence mining without candidate maintenance. *IEEE Transactions on Knowledge and Data Engineering*, 19(8) :1042–1056, 2007.

[WHLS06] Junmei Wang, Wynne Hsu, Mong Li Lee, and Chang Sheng. A Partition-Based Approach to Graph Mining. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, pages 74—-74. IEEE Computer Society, 2006.

[WHLW04] Junmei Wang, Wynne Hsu, Mong-Li Lee, and Jason Tsong-Li Wang. FlowMiner : Finding Flow Patterns in Spatio-Temporal Databases. In *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 14–21. IEEE Computer Society, 2004.

[WM03] Takashi Washio and Hiroshi Motoda. State of the art of graph-based data mining. *SIGKDD Exploration Newsletter*, 5(1) :59–68, 2003.

[WMM05] Takashi Washio, Yuki Mitsunaga, and Hiroshi Motoda. Mining quantitative frequent itemsets using adaptive density-based subspace clustering. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 793–796. IEEE Computer Society, 2005.

[WS78] W.H. Wischmeier and D.D. Smith. Predicting rainfall erosion losses - A guide to conservation planning. 1978.

[WWM+99] Ian H Witten, Ian H Witten, Alistair Moffat, Timothy C Bell, Timothy C Bell, and Timothy C Bell. *Managing gigabytes : compressing and indexing documents and images*. Morgan Kaufmann, 1999.

[YCP+13] Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. Semantic trajectories : Mobility data computation and annotation. *ACM Transactions on Intelligent Systems and Technology*, 4(3) :49, 2013.

[YH02] Xifeng Yan and Jiawei Han. gSpan : Graph-Bases Substructure Pattern Mining. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, volume 3, pages 721–724. IEEE Computer Society, 2002.

[YH03] Xifeng Yan and Jiawei Han. CloseGraph. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 6, page 286, New York, New York, USA, 2003. ACM Press.

[YHA03] Xifeng Yan, Jiawei Han, and Ramin Afshar. Clospan : Mining : Closed sequential patterns in large datasets. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 166–177. SIAM, 2003.

[YL06] Unil Yun and John J Leggett. Wspan : Weighted sequential pattern mining in large sequence databases. In *Intelligent Systems, 2006 3rd International IEEE Conference on*, pages 512–517. IEEE Computer Society, 2006.

[YPM05] Hui Yang, Srinivasan Parthasarathy, and Sameep Mehta. A generalized framework for mining spatio-temporal patterns in scientific data. In Robert Grossman, Roberto J Bayardo, and Kristin P Bennett, editors, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 716–721. ACM, 2005.

[YS06] Jin Soung Yoo and Shashi Shekhar. A Joinless Approach for Mining Spatial Colocation Patterns. *IEEE Transactions on Knowledge and Data Engineering*, 18(10) :1323–1337, 2006.

[Yu16] Wenhao Yu. Spatial co-location pattern mining for location-based services in road networks. *Expert Systems with Applications*, 46 :324–335, 2016.

[YWY07] Junsong Yuan, Ying Wu, and Ming Yang. Discovery of collocation patterns : from visual words to visual phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE Computer Society, 2007.

[YZC12] Junfu Yin, Zhigang Zheng, and Longbing Cao. Uspan : an efficient algorithm for mining high utility sequential patterns. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 660–668. ACM, 2012.

[Zak00a] Mohammed J Zaki. Sequence mining in categorical domains : incorporating constraints. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 422–429. ACM, 2000.

[Zak00b] Mohammed Javeed Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3) :372–390, 2000.

[Zak01] Mohammed J Zaki. Spade : An efficient algorithm for mining frequent sequences. *Machine learning*, 42(1-2) :31–60, 2001.

[ZH02] Mohammed Javeed Zaki and Ching-Jiu Hsiao. CHARM : An Efficient Algorithm for Closed Itemset Mining. In Robert L Grossman, Jiawei Han, Vipin Kumar, Heikki Mannila, and Rajeev Motwani, editors, *Proceedings of the SIAM International Conference on Data Mining (SDM)*. SIAM, 2002.

[ZLY09] Shijie Zhang, Shirong Li, and Jiong Yang. Gaddi : distance index based subgraph matching in biological networks. In *Proceedings of the International Conference on Extending Database Technology : Advances in Database Technology (EDBT)*, pages 192–203. ACM, 2009.

[ZM11] Chunjie Zhou and Xiaofeng Meng. Sts : complex spatio-temporal sequence mining in flickr. In *International Conference on Database Systems for Advanced Applications*, pages 208–223. Springer, 2011.

[ZN14] Albrecht Zimmermann and Siegfried Nijssen. Supervised pattern mining and applications to classification. In *Frequent pattern mining*, pages 425–442. Springer, 2014.

[ZPO+97] Mohammed Javeed Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, Wei Li, et al. New algorithms for fast discovery of association rules. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 97, pages 283–286, 1997.

[ZPOL97] Mohammed Javeed Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, and Wei Li. New Algorithms for Fast Discovery of Association Rules. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 283–286, 1997.

[ZW04] Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise : A quantitative study. *Artificial intelligence review*, 22(3) :177–210, 2004.